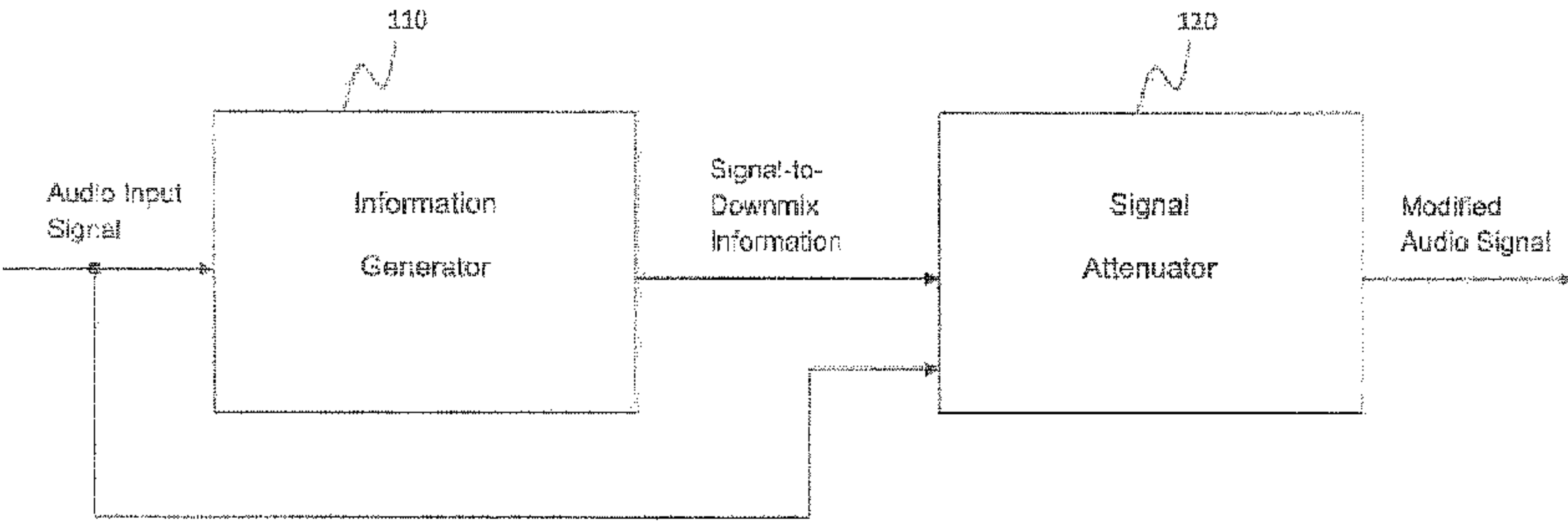




(86) Date de dépôt PCT/PCT Filing Date: 2014/04/07 (87) Date publication PCT/PCT Publication Date: 2014/10/16 (45) Date de délivrance/Issue Date: 2019/08/20 (85) Entrée phase nationale/National Entry: 2015/10/06 (86) N° demande PCT/PCT Application No.: EP 2014/056917 (87) N° publication PCT/PCT Publication No.: 2014/166863 (30) Priorités/Priorities: 2013/04/12 (EP13163621.9); 2013/08/28 (EP13182103.5)	(51) Cl.Int./Int.Cl. <i>H04S 7/00</i> (2006.01), <i>H04S 3/00</i> (2006.01) (72) Inventeurs/Inventors: UHLE, CHRISTIAN, DE; PROKEIN, PETER, DE; HELLMUTH, OLIVER, DE; SCHARRER, SEBASTIAN, DE; HABETS, EMANUEL, DE (73) Propriétaire/Owner: FRAUNHOFER-GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG E.V., DE (74) Agent: PERRY + CURRIER
---	--

(54) Titre : APPAREIL ET PROCEDE DE MISE A L'ECHELLE DE SIGNAL CENTRALE ET AMELIORATION STEREOPHONIQUE BASEE SUR UN RAPPORT DE MIXAGE REDUCTEUR PAR RAPPORT A UN SIGNAL

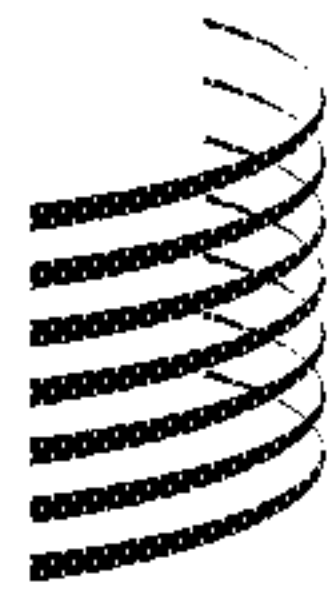
(54) Title: APPARATUS AND METHOD FOR CENTER SIGNAL SCALING AND STEREOPHONIC ENHANCEMENT BASED ON A SIGNAL-TO-DOWNMIX RATIO



(57) **Abrégé/Abstract:**
An apparatus for generating a modified audio signal comprising two or more modified audio channels from an audio input signal comprising two or more audio input channels is provided. The apparatus comprises an information generator (110) for generating signal- to-downmix information. The information generator (110) is adapted to generate signal information by combining a spectral value of each of the two or more audio input channels in a first way. Moreover, the information generator (110) is adapted to generate down mix information by combining the spectral value of each of the two or more audio input channels in a second way being different from the first way. Furthermore, the information generator (110) is adapted to combine the signal information and the downmix information to obtain signal-to-downmix information. Moreover, the apparatus comprises a signal attenuator (120) for attenuating the two or more audio input channels depending on the signal-to-downmix information to obtain the two or more modified audio channels.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
16 October 2014 (16.10.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/166863 A1

(51) International Patent Classification:
H04S 7/00 (2006.01) *H04S 3/00* (2006.01)

(74) Agents: ZINKLER, Franz et al.; Postfach 246, 82043 Pullach (DE).

(21) International Application Number:
PCT/EP2014/056917

(81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(22) International Filing Date:
7 April 2014 (07.04.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
13163621.9 12 April 2013 (12.04.2013) EP
13182103.5 28 August 2013 (28.08.2013) EP

(71) Applicants: FRAUNHOFER-GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E.V. [DE/DE]; Hansastraße 27c, 80686 München (DE). FRIEDRICH-ALEXANDER-UNIVERSITÄT ERLANGEN-NÜRNBERG [DE/DE]; Schlossplatz 4, 91054 Erlangen (DE).

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(72) Inventors: UHLE, Christian; Hoher Rain 28, 92289 Ursensollen (DE). PROKEIN, Peter; Walburgastraße 13, 91056 Erlangen (DE). HELLMUTH, Oliver; Geschwister-Voemel-Weg 60, 91052 Erlangen (DE). SCHARRER, Sebastian; Arzbergweg 39b, 91217 Hersbruck (DE). HABETS, Emanuel; Schwedenstraße 13, 91080 Spardorf (DE).

Published:

— with international search report (Art. 21(3))

(54) Title: APPARATUS AND METHOD FOR CENTER SIGNAL SCALING AND STEREOPHONIC ENHANCEMENT BASED ON A SIGNAL-TO-DOWNMIX RATIO

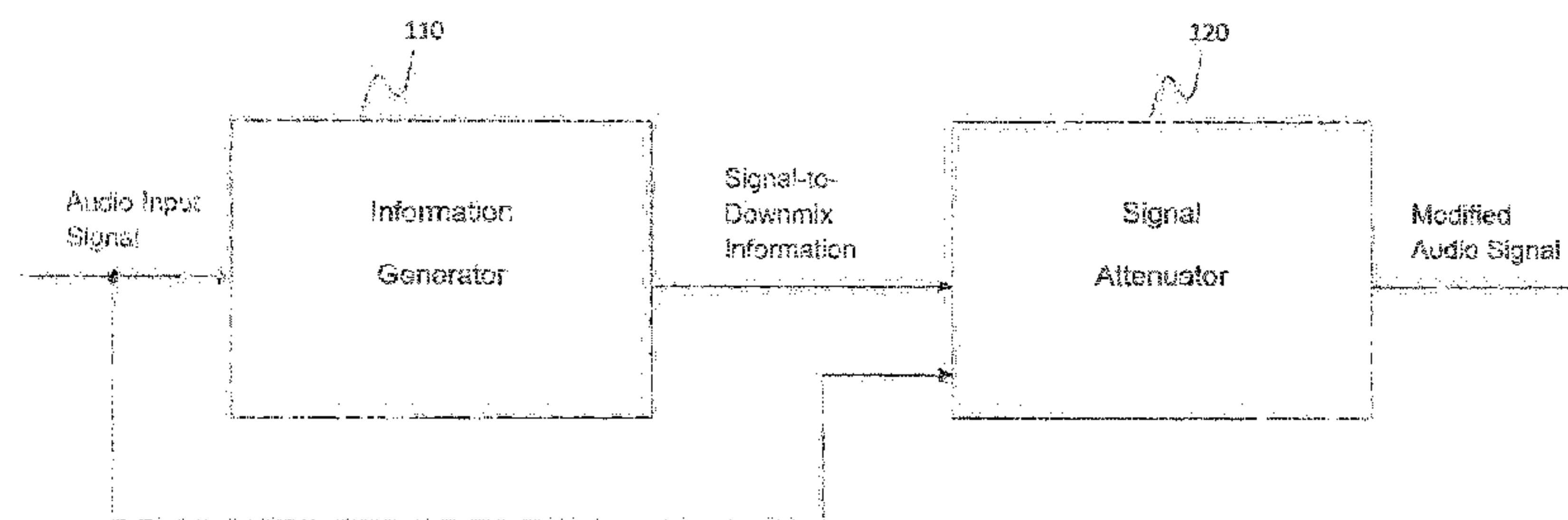


FIG 1

(57) Abstract: An apparatus for generating a modified audio signal comprising two or more modified audio channels from an audio input signal comprising two or more audio input channels is provided. The apparatus comprises an information generator (110) for generating signal- to-downmix information. The information generator (110) is adapted to generate signal information by combining a spectral value of each of the two or more audio input channels in a first way. Moreover, the information generator (110) is adapted to generate down mix information by combining the spectral value of each of the two or more audio input channels in a second way being different from the first way. Furthermore, the information generator (110) is adapted to combine the signal information and the downmix information to obtain signal-to-downmix information. Moreover, the apparatus comprises a signal attenuator (120) for attenuating the two or more audio input channels depending on the signal-to-downmix information to obtain the two or more modified audio channels.

WO 2014/166863 A1

Apparatus and Method for Center Signal Scaling and Stereophonic Enhancement based on a Signal-to-Downmix Ratio

Description

5

The present invention relates to audio signal processing, and, in particular, to a center signal scaling and stereophonic enhancement based on the signal-to-downmix ratio.

- 10 Audio signals are in general a mixture of direct sounds and ambient (or diffuse) sounds. Direct signals are emitted by sound sources, e.g. a musical instrument, a vocalist or a loudspeaker, and arrive on the shortest possible path at the receiver, e.g. the listener's ear or a microphone. When listening to a direct sound, it is perceived as coming from the direction of the sound source. The relevant auditory cues for the localization and for other
- 15 spatial sound properties are interaural level difference (ILD), interaural time difference (ITD) and interaural coherence. Direct sound waves evoking identical ILD and ITD are perceived as coming from the same direction. In the absence of ambient sound, the signals reaching the left and the right ear or any other set of spaced sensors are coherent.
- 20 Ambient sounds, in contrast, are emitted by many spaced sound sources or sound reflecting boundaries contributing to the same sound. When a sound wave reaches a wall in a room, a portion of it is reflected, and the superposition of all reflections in a room, the reverberation, is a prominent example for ambient sounds. Other examples are applause, babble noise and wind noise. Ambient sounds are perceived as being diffuse, not
- 25 locatable, and evoke an impression of envelopment (of being "immersed in sound") by the listener. When capturing an ambient sound field using a set of spaced sensors, the recorded signals are at least partially incoherent.

Related prior art on separation, decomposition or scaling is either based on panning

30 information, i.e. inter-channel level differences (ICLD) and inter-channel time differences (ICTD), or based on signal characteristics of direct and of ambient sounds. Methods taking advantage of ICLD in two-channel stereophonic recordings are the upmix method described in [7], the Azimuth Discrimination and Resynthesis (ADRes) algorithm [8], the upmix from two-channel input signals to three channels proposed by Vickers [9], and the

35 center signal extraction described in [10].

The Degenerate Unmixing Estimation Technique (DUET) [11, 12] is based on clustering the time-frequency bins into sets with similar ICLD and ICTD. A restriction of the original

method is that the maximum frequency which can be processed equals half the speed of sound over maximum microphone spacing (due to ambiguities in the ICTD estimation) which has been addressed in [13]. The performance of the method decreases when sources overlap in the time-frequency domain and when the reverberation increases.

5 Other methods based on ICLD and ICTD are the Modified ADress algorithm [14], which extends ADress algorithm [8] for the processing of spaced microphone recordings, the method based on time-frequency correlation (AD-TIFCORR) [15] for time-delayed mixtures, the Direction Estimation of Mixing Matrix (DEMIX) for anechoic mixtures [16], which includes a confidence measure that only one source is active at a particular time-
10 frequency bin, the Model-based Expectation-Maximization Source Separation and Localization (MESSL) [17], and methods mimicking the binaural human hearing mechanism as in e.g. [18, 19].

Despite the methods for Blind Source Separation (BSS) using spatial cues of direct signal
15 components mentioned above, also the extraction and attenuation of ambient signals are related to the presented method. Methods based on the inter-channel coherence (ICC) in two-channel signals are described in [22, 7, 23]. The application of adaptive filtering has been proposed in [24], with the rationale that direct signals can be predicted across channels whereas diffuse sounds are obtained from the prediction error.

20

A method for upmixing of two-channel stereophonic signals based on multichannel Wiener filtering estimates both, the ICLD of direct sounds and the power spectral densities (PSD) of the direct and ambient signal components [25].

25 Approaches to the extraction of ambient signals from single channel recordings include the use of Non-Negative Matrix Factorization of a time-frequency representation of the input signal, where the ambient signal is obtained from the residual of that approximation [26], low-level feature extraction and supervised learning [27], and the estimation of the impulse response of a reverberant system and inverse filtering in the frequency domain
30 [28].

The object of the present invention is to provide improved concepts for audio signal processing. The object of the present invention is solved by an apparatus according to claim 1, by a system according to claim 14, by a method according to claim 15 and by a
35 computer program according to claim 16.

An apparatus for generating a modified audio signal comprising two or more modified audio channels from an audio input signal comprising two or more audio input channels is

provided. The apparatus comprises an information generator for generating signal-to-downmix information. The information generator is adapted to generate signal information by combining a spectral value of each of the two or more audio input channels in a first way. Moreover, the information generator is adapted to generate downmix information by
5 combining the spectral value of each of the two or more audio input channels in a second way being different from the first way. Furthermore, the information generator is adapted to combine the signal information and the downmix information to obtain signal-to-downmix information. Moreover, the apparatus comprises a signal attenuator for attenuating the two or more audio input channels depending on the signal-to-downmix
10 information to obtain the two or more modified audio channels.

In a particular embodiment, the apparatus may, for example, be adapted to generate a modified audio signal comprising three or more modified audio channels from an audio input signal comprising three or more audio input channels.

15 In an embodiment, the number of the modified audio channels is equal to or smaller than the number of the audio input channels, or wherein the number of the modified audio channels is smaller than the number of the audio input channels. For example, according to a particular embodiment, the apparatus may be adapted to generate a modified audio
20 signal comprising two or more modified audio channels from an audio input signal comprising two or more audio input channels, wherein the number of the modified audio channels is equal to the number of the audio input channels.

Embodiments provide new concepts for scaling the level of the virtual center in audio
25 signals is proposed. The input signals are processed in the time-frequency domain such that direct sound components having approximately equal energy in all channels are amplified or attenuated. The real-valued spectral weights are obtained from the ratio of the sum of the power spectral densities of all input channel signals and the power spectral density of the sum signal. Applications of the presented concepts are upmixing two-
30 channel stereophonic recordings for its reproduction using surround sound set-ups, stereophonic enhancement, dialogue enhancement, and as preprocessing for semantic audio analysis.

Embodiments provide new concepts for amplifying or attenuating the center signal in an
35 audio signal. In contrast to previous concepts, both, lateral displacement and diffuseness of the signal components are taken into account. Furthermore, the use of semantically meaningful parameters is discussed in order to support the user when implementations of the concepts are employed.

Some embodiments focus on center signal scaling, i.e. the amplification or attenuation of center signals in audio recordings. The center signal is, e.g., defined here as the sum of all direct signal components having approximately equal intensity in all channels and negligible time differences between the channels.

Various applications of audio signal processing and reproduction benefit from center signal scaling, e.g. upmixing, dialogue enhancement, and semantic audio analysis.

Upmixing refers to the process of creating an output signal given an input signal with less channels. Its main application is the reproduction of two-channel signals using surround sound setups as for example specified in [1]. Research on the subjective quality of spatial audio [2] indicates that locatedness [3], localization and width are prominent descriptive attributes of sound. Results of a subjective assessment of 2-to-5 upmixing algorithms [4] showed that the use of an additional center loudspeaker can narrow the stereophonic image. The presented work is motivated by the assumption that locatedness, localization and width can be preserved or even improved when the additional center loudspeaker reproduces mainly *direct* signal components which are panned to the center, and when these signal components are attenuated in the off-center loudspeaker signals.

Dialogue enhancement refers to the improvement of speech intelligibility, e.g. in broadcast and movie sound, and is often desired when background sounds are too loud relative to the dialogue [5]. This applies in particular to persons who are hard of hearing, non-native listeners, in noisy environments or when the binaural masking level difference is reduced due to narrow loudspeaker placement. The concepts method can be applied for processing input signals where the dialogue is panned to the center in order to attenuate background sounds and thereby enabling better speech intelligibility.

Semantic Audio Analysis (or Audio Content Analysis) comprises processes for deducing meaningful descriptors from audio signals, e.g. beat tracking or transcription of the leading melody. The performance of the computational methods is often deteriorated when the sounds of interest are embedded in background sounds, see e.g. [6]. Since it is common practice in audio production that sound sources of interest (e.g. leading instruments and singers) are panned to the center, center extraction can be applied as a pre-processing step for attenuating background sounds and reverberation.

According to an embodiment, the information generator may be configured to combine the signal information and the downmix information so that the signal-to-downmix information indicates a ratio of the signal information to the downmix information.

- 5 In an embodiment, the information generator may be configured to process the spectral value of each of the two or more audio input channels to obtain two or more processed values, and wherein the information generator may be configured to combine the two or more processed values to obtain the signal information. Moreover, the information generator may be configured to combine the spectral value of each of the two or more
10 audio input channels to obtain a combined value, and wherein the information generator may be configured to process the combined value to obtain the downmix information.

According to an embodiment, the information generator may be configured to process the spectral value of each of the two or more audio input channels by multiplying said spectral
15 value by the complex conjugate of said spectral value to obtain an auto power spectral density of said spectral value for each of the two or more audio input channels.

In an embodiment, the information generator may be configured to process the combined value by determining a power spectral density of the combined value.

20

According to an embodiment, the information generator may be configured to generate the signal information $s(m, k, \beta)$ according to the formula:

$$s(m, k, \beta) = \sum_{i=1}^N \Phi_{i,i}(m, k)^{\beta},$$

25

wherein N indicates the number of audio input channels of the audio input signal, wherein $\Phi_{i,i}(m, k)$ indicates the auto power spectral density of the spectral value of the i -th audio signal channel, wherein β is a real number with $\beta > 0$, wherein m indicates a time index, and wherein k indicates a frequency index. For example, according to a particular
30 embodiment $\beta \geq 1$.

In an embodiment, the information generator may be configured to determine the signal-to-downmix ratio as the signal-to-downmix information according to the formula $R(m, k, \beta)$

$$R(m, k, \beta) = \left(\frac{\sum_{i=1}^N \Phi_{i,i}(m, k)^{\beta}}{\Phi_d(m, k)^{\beta}} \right)^{\frac{1}{2\beta-1}},$$

35

wherein $\Phi_d(m, k)$ indicates the power spectral density of the combined value, and wherein $\Phi_d(m, k)^\beta$ is the downmix information.

- 5 According to an embodiment, the information generator may be configured to generate the signal information $\Phi_1(m, k)$ according to the formula

$$\Phi_1(m, k) = \mathcal{E}\{\mathbf{W}\mathbf{X}(m, k)(\mathbf{W}\mathbf{X}(m, k))^H\},$$

- 10 wherein the information generator is configured to generate the downmix information $\Phi_2(m, k)$ according to the formula

$$\Phi_2(m, k) = \mathcal{E}\{\mathbf{V}\mathbf{X}(m, k)(\mathbf{V}\mathbf{X}(m, k))^H\}, \text{ and}$$

- 15 wherein the information generator is configured to generate the signal-to-downmix ratio as the signal-to-downmix information $R_g(m, k, \beta)$ according to the formula

$$R_g(m, k, \beta) = \left(\frac{\text{tr}\{\Phi_1(m, k)^\beta\}}{\text{tr}\{\Phi_2(m, k)^\beta\}} \right)^{\frac{1}{2\beta-1}}$$

- 20 wherein $\mathbf{X}(m, k)$ indicates the audio input signal, wherein

$$\mathbf{X}(m, k) = [X_1(m, k) \cdots X_N(m, k)]^T$$

- wherein N indicates the number of audio input channels of the audio input signal, wherein
 25 m indicates a time index, and wherein k indicates a frequency index, wherein $X_1(m, k)$ indicates the first audio input channel, wherein $X_N(m, k)$ indicates the N -th audio input channel, wherein $\mathbf{V}$ indicates a matrix or a vector, wherein $\mathbf{W}$ indicates a matrix or a vector, wherein ^{H} indicates the conjugate transpose of a matrix or a vector, wherein $\mathcal{E}\{\cdot\}$ is an expectation operation, wherein β is a real number with $\beta > 0$, and wherein $\text{tr}\{\cdot\}$ is the
 30 trace of a matrix. For example, according to a particular embodiment $\beta \geq 1$.

In an embodiment, $\mathbf{V}$ may be a row vector of length N whose elements are equal to one and $\mathbf{W}$ may be the identity matrix of size $N \times N$.

- 35 According to an embodiment, $\mathbf{V} = [1, 1]$, wherein $\mathbf{W} = [1, -1]$ and wherein $N = 2$.

In an embodiment, the signal attenuator may be adapted to attenuate the two or more audio input channels depending on a gain function $G(m, k)$ according to the formula

$$5 \quad Y(m, k) = G(m, k)X(m, k),$$

wherein the gain function $G(m, k)$ depends on the signal-to-downmix information, and wherein the gain function $G(m, k)$ is a monotonically increasing function of the signal-to-downmix information or a monotonically decreasing function of the signal-to-downmix information,

wherein $X(m, k)$ indicates the audio input signal, wherein $Y(m, k)$ indicates the modified audio signal, wherein m indicates a time index, and wherein k indicates a frequency index.

15 According to an embodiment, the gain function $G(m, k)$ may be a first function $G_{c1}(m, k, \beta, \gamma)$, a second function $G_{c2}(m, k, \beta, \gamma)$, a third function $G_{s1}(m, k, \beta, \gamma)$ or a fourth function $G_{s2}(m, k, \beta, \gamma)$,

wherein

$$20 \quad G_{c1}(m, k, \beta, \gamma) = (1 + R_{\min} - R(m, k, \beta))^{\gamma},$$

wherein

$$25 \quad G_{c2}(m, k, \beta, \gamma) = \left(\frac{R_{\min}}{R(m, k, \beta)} \right)^{\gamma},$$

wherein

$$30 \quad G_{s1}(m, k, \beta, \gamma) = R(m, k, \beta)^{\gamma},$$

wherein

$$35 \quad G_{s2}(m, k, \beta, \gamma) = \left(1 + R_{\min} - \frac{R_{\min}}{R(m, k, \beta)} \right)^{\gamma},$$

wherein β is a real number with $\beta > 0$,

wherein γ is a real number with $\gamma > 0$, and

wherein $R_{\min}$ indicates the minimum of R .

5

Moreover, a system is provided. The system comprises a phase compensator for generating a phase-compensated audio signal comprising two or more phase-compensated audio channels from an unprocessed audio signal comprising two or more unprocessed audio channels. Furthermore, the system comprises an apparatus according to one of the above-described embodiments for receiving the phase compensated audio signal as an audio input signal and for generating a modified audio signal comprising two or more modified audio channels from the audio input signal comprising the two or more phase-compensated audio channels as two or more audio input channels. One of the two or more unprocessed audio channels is a reference channel. The phase compensator is adapted to estimate for each unprocessed audio channel of the two or more unprocessed audio channels which is not the reference channel a phase transfer function between said unprocessed audio channel and the reference channel. Moreover, the phase compensator is adapted to generate the phase-compensated audio signal by modifying each unprocessed audio channel of the unprocessed audio channels which is not the reference channel depending on the phase transfer function of said unprocessed audio channel.

20

Furthermore, a method for generating a modified audio signal comprising two or more modified audio channels from an audio input signal comprising two or more audio input channels is provided. The method comprises:

25

- Generating signal information by combining a spectral value of each of the two or more audio input channels in a first way.
- Generating downmix information by combining the spectral value of each of the two or more audio input channels in a second way being different from the first way.
- Generating signal-to-downmix information by combining the signal information and the downmix information. And:
- Attenuating the two or more audio input channels depending on the signal-to-downmix information to obtain the two or more modified audio channels.

35

Moreover, a computer program for implementing the above-described method when being executed on a computer or signal attenuator is provided.

In the following, embodiments of the present invention are described in more detail with
5 reference to the figures, in which:

- Fig. 1 illustrates an apparatus according to an embodiment,
- Fig. 2 illustrates the signal-to-downmix ratio as function of the inter-channel level
10 differences and as a function of the inter-channel coherence according to an embodiment,
- Fig. 3 illustrates spectral weights as a function of the inter-channel coherence *and*
15 of the inter-channel level differences according to an embodiment,
- Fig. 4 illustrates spectral weights as a function of the inter-channel coherence and
of the inter-channel level differences according to another embodiment,
- Fig. 5 illustrates spectral weights as a function of the inter-channel coherence *and*
20 of the inter-channel level differences according to a further embodiment,
- Fig. 6a-e illustrate spectrograms the direct source signals and the left and right
channel signals of the mixture signal,
- 25 Fig. 7 illustrates the input signal and the output signal for the center signal
extraction according to an embodiment,
- Fig. 8 illustrates the spectrograms of the output signal according to an
embodiment,
- 30 Fig. 9 illustrates the input signal and the output signal for the center signal
attenuation according to another embodiment,
- Fig. 10 illustrates the spectrograms of the output signal according to an
35 embodiment,
- Fig. 11a-d illustrate two speech signals which have been mixed to obtain input signals
with and without inter-channel time differences,

Fig. 12a-c illustrate the spectral weights computed from a gain function according to an embodiment, and

5 Fig. 13 illustrates a system according to an embodiment.

Fig. 1 illustrates an apparatus for generating a modified audio signal comprising two or more modified audio channels from an audio input signal comprising two or more audio
10 input channels according to an embodiment.

The apparatus comprises an information generator 110 for generating signal-to-downmix information.

15 The information generator 110 is adapted to generate signal information by combining a spectral value of each of the two or more audio input channels in a first way. Moreover, the information generator 110 is adapted to generate downmix information by combining the spectral value of each of the two or more audio input channels in a second way being different from the first way.

20

Furthermore, the information generator 110 is adapted to combine the signal information and the downmix information to obtain signal-to-downmix information. For example, the signal-to-downmix information may be a signal-to-downmix ratio, e.g., a signal-to-downmix value.

25

Moreover, the apparatus comprises a signal attenuator 120 for attenuating the two or more audio input channels depending on the signal-to-downmix information to obtain the two or more modified audio channels.

30 According to an embodiment, the information generator may be configured to combine the signal information and the downmix information so that the signal-to-downmix information indicates a ratio of the signal information to the downmix information. For example, the signal information may be a first value and the downmix information may be a second value and the signal-to-downmix information indicates a ratio of the signal value to the
35 downmix value. For example, the signal-to-downmix information may be the first value divided by the second value. Or, for example, if the first value and the second value are logarithmic values, the signal-to-downmix information may be the difference between the first value and the second value.

In the following, the underlying signal model and the concepts are described and analyzed for the case of input signal featuring amplitude difference stereophony.

- 5 The rationale is to compute and apply real-valued spectral weights as a function of the diffuseness and the lateral position of direct sources. The processing as demonstrated here is applied in the STFT domain, yet it is not restricted to a particular filterbank. The N channel input signal is denoted by

$$10 \quad \mathbf{x}[n] = [x_1[n] \cdots x_N[n]]^T. \quad (1)$$

where n denotes the discrete time index. The input signal is assumed to be an additive mixture of direct signals $s_i[n]$ and ambient sounds $a_l[n]$,

$$15 \quad x_l[n] = \sum_{i=1}^K d_{i,l}[n] * s_i[n] + a_l[n], l = 1, \dots, N \quad (2)$$

where P is the number of sound sources, $d_{i,l}[n]$ denote the impulse responses of the direct paths of the i -th source into the l -th channel of length $L_{i,l}$ samples, and the ambient signal components are mutually uncorrelated or weakly correlated. In the following
20 description it is assumed that the signal model corresponds to amplitude difference stereophony, i.e. $L_{i,l} = 1, \forall i, l$.

The time-frequency domain representation of $\mathbf{x}[n]$ is given by

$$25 \quad \mathbf{X}(m, k) = [X_1(m, k) \cdots X_N(m, k)]^T, \quad (3)$$

with time index m and frequency index k . The output signals are denoted by

$$30 \quad \mathbf{Y}(m, k) = [Y_1(m, k) \cdots Y_N(m, k)]^T, \quad (4)$$

and are obtained by means of spectral weighting

$$\mathbf{Y}(m, k) = G(m, k) \mathbf{X}(m, k), \quad (5)$$

with real-valued weights $G(m, k)$. Time domain output signals are computed by applying the inverse processing of the filterbank. For the computation of the spectral weights, the sum signal, thereafter denoted as the downmix signal, is computed as

$$X_d(m, k) = \sum_{i=1}^N X_i(m, k) , \quad (6)$$

The matrix of PSD of the input signal, comprising estimates of the (auto-)PSD on the main diagonal, while off-diagonal elements are estimates of the cross-PSD, is given by

$$\Phi_{i,l}(m, k) = \mathcal{E} \{ X_i(m, k) X_l^*(m, k) \} , \quad i, l = 1 \dots N , \quad (7)$$

where X^* denotes the complex conjugate of X , and $\mathcal{E} \{ \cdot \}$ is the expectation operation with respect to the time dimension. In the presented simulations the expectation values are estimated using single-pole recursive averaging,

$$\Phi_{i,l}(m, k) = \alpha X_i(m, k) X_l^*(m, k) + (1 - \alpha) \Phi_{i,l}(m - 1, k) , \quad (8)$$

where the filter coefficient α determines the integration time. Furthermore, the quantity $R(m, k; \beta)$ is defined as

$$R(m, k, \beta) = \left(\frac{\sum_{i=1}^N \Phi_{i,i}(m, k)^\beta}{\Phi_d(m, k)^\beta} \right)^{\frac{1}{2\beta-1}} , \quad (9)$$

where $\Phi_d(m, k)$ is the PSD of the downmix signal and β is a parameter which will be addressed in the following. The quantity $R(m, k; 1)$ is the signal-to-downmix ratio (SDR), i.e. the ratio of the total PSD and the PSD of the downmix signal. The power to $\frac{1}{2\beta-1}$ ensures that the range of $R(m, k; \beta)$ is independent of β .

The information generator 110 may be configured to determine the signal-to-downmix ratio according to Equation (9).

According to Equation (9) the signal information $s(m, k, \beta)$ that may be determined by the information generator 110 is defined as

$$s(m, k, \beta) = \sum_{i=1}^N \Phi_{i,i}(m, k)^{\beta}.$$

As can be seen above, $\Phi_{i,i}(m, k)$ is defined as $\Phi_{i,i}(m, k) = \mathcal{E} \{ X_i(m, k) X_i^*(m, k) \}$. Thus, to determine the signal information $s(m, k, \beta)$, the spectral value $X_i(m, k)$ of each of the two or more audio input channels is processed to obtain the processed value $\Phi_{i,i}(m, k)^{\beta}$ for each of the two or more audio input channels, and the obtained processed values $\Phi_{i,i}(m, k)^{\beta}$ are then combined, e.g., as in Equation (9) by summing up the obtained processed values $\Phi_{i,i}(m, k)^{\beta}$.

Thus, the information generator 110 may be configured to process the spectral value $X_i(m, k)$ of each of the two or more audio input channels to obtain two or more processed values $\Phi_{i,i}(m, k)^{\beta}$, and the information generator 110 may be configured to combine the two or more processed values to obtain the signal information $s(m, k, \beta)$. In more general, the information generator 110 is adapted to generate signal information $s(m, k, \beta)$ by combining a spectral value $X_i(m, k)$ of each of the two or more audio input channels in a first way.

Moreover, according to Equation (9) the downmix information $d(m, k, \beta)$ that may be determined by the information generator 110 is defined as

$$d(m, k, \beta) = \Phi_d(m, k)^{\beta}.$$

To form $\Phi_d(m, k)$, at first $X_d(m, k)$ is formed according to the above Equation (6):

$$X_d(m, k) = \sum_{i=1}^N X_i(m, k).$$

As can be seen, at first, the spectral value $X_i(m, k)$ of each of the two or more audio input channels is combined to obtain a combined value $X_d(m, k)$, e.g., as in Equation (6), by summing up the spectral value $X_i(m, k)$ of each of the two or more audio input channels.

Then, to obtain $\Phi_d(m, k)$, the power spectral density of $X_d(m, k)$ is formed, e.g., according to

$$\Phi_d(m, k) = \mathcal{E} \{ X_d(m, k) X_d^*(m, k) \},$$

and then, $\Phi_d(m, k)^\beta$ may be determined. More generally speaking, the obtained combined value $X_d(m, k)$ has been processed to obtain the downmix information $d(m, k, \beta) = \Phi_d(m, k)^\beta$.

5 Thus, the information generator 110 may be configured to combine the spectral value $X_i(m, k)$ of each of the two or more audio input channels to obtain a combined value, and the information generator 110 may be configured to process the combined value to obtain the downmix information $d(m, k, \beta)$. In more general, the information generator 110 is adapted to generate downmix information $d(m, k, \beta)$ by combining the spectral value
10 $X_i(m, k)$ of each of the two or more audio input channels in a second way. The way, how the downmix information is generated ("second way") differs from the way, how the signal information is generated ("first way") and thus, the second way is different from the first way.

15 The information generator 110 is adapted to generate signal information by combining a spectral value of each of the two or more audio input channels in a first way. Moreover, the information generator 110 is adapted to generate downmix information by combining the spectral value of each of the two or more audio input channels in a second way being different from the first way.

20

Fig. 2, upper plot illustrates the signal-to-downmix ratio $R(m, k; 1)$ for $N=2$ as function of the ICLD $\Theta(m, k)$, shown for $\Psi(m, k) \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Fig.2, lower plot illustrates the signal-to-downmix ratio $R(m, k; 1)$ for $N=2$ as function of ICC $\Psi(m, k)$ and ICLD $\Theta(m, k)$ in color-coded 2D-plot.

25

In particular, Fig. 2 illustrates the SDR for $N = 2$ as a function of ICC $\Psi(m, k)$ and ICLD $\Theta(m, k)$, with

$$\Psi(m, k) = \frac{|\Phi_{1,2}(m, k)|}{\sqrt{\Phi_{1,1}(m, k)\Phi_{2,2}(m, k)}}, \quad (10)$$

30

and

$$\Theta(m, k) = \frac{\Phi_{1,1}(m, k)}{\Phi_{2,2}(m, k)}. \quad (11)$$

35 Fig. 2 shows that the SDR has the following properties:

1. It is monotonically related to both, $\Psi(m, k)$ and $|\log \Theta(m, k)|$.
2. For diffuse input signals, i.e. $\Psi(m, k) = 0$, the SDR assumes its maximum value,
 $R(m, k; 1) = 1$.
3. For direct sounds panned to the center, i.e. $\Theta(m, k) = 1$, the SDR assumes its
 minimum value $R_{\min}$, where $R_{\min} = 0.5$ for $N=2$.

Due to these properties, appropriate spectral weights for center signal scaling can be computed from the SDR by using monotonically *decreasing* functions for the *extraction* of center signals and monotonically *increasing* functions for the *attenuation* of center signals.

For the extraction of a center signal, appropriate functions of $R(m, k; \beta)$ are, for example,

$$G_{c1}(m, k, \beta, \gamma) = (1 + R_{\min} - R(m, k, \beta))^{\gamma}, \quad (12)$$

and

$$G_{c2}(m, k, \beta, \gamma) = \left(\frac{R_{\min}}{R(m, k, \beta)} \right)^{\gamma}. \quad (13)$$

where a parameter γ for controlling the maximum attenuation is introduced.

For the attenuation of the center signal, appropriate functions of $R(m, k; \beta)$ are, for example,

$$G_{s1}(m, k, \beta, \gamma) = R(m, k, \beta)^{\gamma}, \quad (14)$$

and

$$G_{s2}(m, k, \beta, \gamma) = \left(1 + R_{\min} - \frac{R_{\min}}{R(m, k, \beta)} \right)^{\gamma}, \quad (15)$$

Fig. 3 and 4 illustrate the gain functions (13) and (15), respectively, for $\beta = 1$, $\gamma = 3$. The spectral weights are constant for $\Psi(m, k) = 0$. The maximum attenuation is $\gamma \cdot 6\text{dB}$, which also applies to the gain functions (12) and (14).

In particular, Fig. 3 illustrates spectral weights $G_{c2}(m, k, 1, 3)$ in dB as function of ICC $\Psi(m, k)$ and ICLD $\Theta(m, k)$.

5 Moreover, Fig. 4 illustrates spectral weights $G_{s2}(m, k, 1, 3)$ in dB as function of ICC $\Psi(m, k)$ and ICLD $\Theta(m, k)$.

Furthermore, Fig. 5 illustrates spectral weights $G_{c2}(m, k, 2, 3)$ in dB as function of ICC $\Psi(m, k)$ and ICLD $\Theta(m, k)$.

10

The effect of the parameter β is shown in Fig. 5 for the gain function in Equation (13) with $\beta = 2, \gamma = 3$. With larger values for β , the influence of Ψ on the spectral weights decreases whereas the influence of Θ increases. This leads to more leakage of diffuse signal components into the output signal, and to more attenuation of the direct signal components panned off-center, when comparing to the gain function in Fig. 3.

15

Post-processing of spectral weights: Prior to the spectral weighting, the weights $G(m, k, \beta, \gamma)$ can be further processed by means of smoothing operations. Zero phase low-pass filtering along the frequency axis reduces circular convolution artifacts which can occur for example when the zero-padding in the STFT computation is too short or a rectangular synthesis window is applied. Low-pass filtering along the time axis can reduce processing artifacts, especially when the time constant for the PSD estimation is rather small.

20

25 In the following, generalized spectral weights are provided.

More general spectral weights are obtained when rewriting Equation (9) as

$$R_g(m, k, \beta) = \left(\frac{\text{tr}\{\Phi_1(m, k)^\beta\}}{\text{tr}\{\Phi_2(m, k)^\beta\}} \right)^{\frac{1}{2\beta-1}}. \quad (16)$$

30

with

$$\Phi_1(m, k) = \mathcal{E}\{\mathbf{W}\mathbf{X}(m, k)(\mathbf{W}\mathbf{X}(m, k))^H\} \quad (17)$$

35

$$\Phi_2(m, k) = \mathcal{E}\{\mathbf{V}\mathbf{X}(m, k)(\mathbf{V}\mathbf{X}(m, k))^H\} \quad (18)$$

where superscript H denotes the conjugate transpose of a matrix or a vector, and $\mathbf{W}$ and $\mathbf{V}$ are mixing matrices or mixing (row) vectors.

Here, $\Phi_1(m, k)$ may be considered as signal information and $\Phi_2(m, k)$ may be considered as
5 downmix information.

For example, $\Phi_2 = \Phi_d$ when $\mathbf{V}$ is a vector of length N whose elements are equal to one. Equation (16) is equal to (9) when $\mathbf{V}$ is a row vector of length N whose elements are equal to one and $\mathbf{W}$ is the identity matrix of size $N \times N$.

10 The generalized SDR $R_g(m, k, \beta, \mathbf{W}, \mathbf{V})$ covers for example the ratio of the PSD of the side signal and of the PSD of the downmix signal, for $\mathbf{W} = [1, -1]$, $\mathbf{V} = [1, 1]$, and $N = 2$.

$$R(m, k, \beta) = \left(\frac{\Phi_s(m, k)^\beta}{\Phi_d(m, k)^\beta} \right)^{\frac{1}{2\beta-1}} \quad (19)$$

15 where $\Phi_s(m, k)$ is the PSD of the side signal.

According to an embodiment, the information generator 110 is adapted to generate signal information $\Phi_1(m, k)$ by combining a spectral value $X_i(m, k)$ of each of the two or more
20 audio input channels in a first way. Moreover, the information generator 110 is adapted to generate downmix information $\Phi_2(m, k)$ by combining the spectral value $X_i(m, k)$ of each of the two or more audio input channels in a second way being different from the first way.

25 In the following, a more general case of mixing models featuring time-of-arrival stereophony is described.

The derivation of the spectral weights described above relies on the assumption that $L_{i,l} = 1, \forall i, l$, i.e. the direct sound sources are time-aligned between the input channels. When the mixing of the direct source signals is not restricted to amplitude difference
30 stereophony ($L_{i,l} > 1$), for example when recording with spaced microphones, the downmix of the input signal $X_d(m, k)$ is subject to phase cancellation. Phase cancellation in $X_d(m, k)$ leads to increasing SDR values and consequently to the typical comb-filtering artifacts when applying the spectral weighting as described above.

35 The notches of the comb-filter correspond to the frequencies

$$f_n = \frac{o f_s}{2d}$$

for gain functions (12) and (13) and

$$5 \quad f_n = \frac{e f_s}{2d}$$

for gain functions (14) and (15), where f_s is the sampling frequency, o are odd integers, e are even integers, and d is the delay in samples.

- 10 A first approach to solve this problem is to compensate the phase differences resulting from the ICTD prior to the computation of $X_d(m, k)$. Phase difference compensation (PDC) is achieved by estimating the time-variant inter-channel phase transfer function $\hat{P}_i(m, k)$ $\hat{P}_i(m, k) \in [-\pi \pi]$ between the i -th channel and a reference channel denoted by index r ,

$$15 \quad \hat{P}_i(m, k) = \arg X_r(m, k) - \arg X_i(m, k) , \quad i \in [1, \dots, N] \setminus r \quad (20)$$

where the operator $A \setminus B$ denotes set-theoretic difference of set B and set A , and applying a time-variant allpass compensation filter $H_{C,i}(m, k)$ to the i -th channel signal

$$20 \quad \tilde{X}_i(m, k) = H_{C,i}(m, k) X_i(m, k) . \quad (21)$$

where the phase transfer function of $H_{C,i}(m, k)$ is

$$\arg H_{C,i}(m, k) = -\mathcal{E} \{ \hat{P}_i(m, k) \} . \quad (22)$$

25

The expectation value is estimated using single-pole recursive averaging. It should be noted that phase jumps of 2π occurring at frequencies close to the notch frequencies need to be compensated for prior to the recursive averaging.

- 30 The downmix signal is computed according to

$$X_d(m, k) = \sum_{i=1}^N \tilde{X}_i(m, k) , \quad (23)$$

- 35 such that the PDC is only applied for computing X_d and does not affect the phase of the output signal.

Fig. 13 illustrates a system according to an embodiment.

The system comprises a phase compensator 210 for generating a phase-compensated
5 audio signal comprising two or more phase-compensated audio channels from an
unprocessed audio signal comprising two or more unprocessed audio channels.

Furthermore, the system comprises an apparatus 220 according to one of the above-
described embodiments for receiving the phase compensated audio signal as an audio
10 input signal and for generating a modified audio signal comprising two or more modified
audio channels from the audio input signal comprising the two or more phase-
compensated audio channels as two or more audio input channels.

One of the two or more unprocessed audio channels is a reference channel. The phase
15 compensator 210 is adapted to estimate for each unprocessed audio channel of the two
or more unprocessed audio channels which is not the reference channel a phase transfer
function between said unprocessed audio channel and the reference channel. Moreover,
the phase compensator 210 is adapted to generate the phase-compensated audio signal
by modifying each unprocessed audio channel of the unprocessed audio channels which
20 is not the reference channel depending on the phase transfer function of said
unprocessed audio channel.

In the following, intuitive explanations of the control parameters are provided, e.g., a
semantic meaning of control parameters.

25

For the operation of digital audio effects it is advantageous to provide controls with
semantically meaningful parameters. The gain functions (12) - (15) are controlled by the
parameters α , β and γ . Sound engineers and audio engineers are used to time constants,
and specifying α as time constant is intuitive and according to common practice. The
30 effect of the integration time can be experienced best by experimentation. In order to
support the operation of the provided concepts, descriptors for the remaining parameters
are proposed, namely *impact* for γ and *diffuseness* for β .

The parameter *impact* can be best compared with the order of a filter. By analogy to the
35 roll-off in filtering, the maximum attenuation equals $\gamma \cdot 6\text{dB}$, for $N = 2$.

The label *diffuseness* is proposed here to emphasize the fact that then attenuating
panned and diffuse sounds, larger values of β result in more leakage of diffuse sounds. A

nonlinear mapping of the user parameter β_u , e.g. $\beta = \sqrt{\beta_u + 1}$, with $0 \leq \beta_u \leq 10$, is advantageous in a way that it enables a more consistent behavior of the processing as opposed to when modifying β directly (where *consistency* relates to the effect of a change of the parameter on the result throughout the range of the parameter value).

5

In the following, computational complexity and memory requirements are briefly discussed.

The computational complexity and memory requirements scale with the number of bands of the filterbank and depend on the implementation of additional post-processing of the spectral weights. A low-cost implementation of the method can be achieved when setting $\beta = 1$, $\gamma \in \mathbb{N}$, computing spectral weights according to Equation (12) or (14), and when not applying the PDC filter. The computation of the SDR uses only one cost intensive nonlinear functions per sub-band when $\beta \in \mathbb{N}$. For $\beta = 1$, only two buffers for the PSD estimation are required, whereas methods making explicit use of the ICC, e.g. [7, 10, 20, 21, 23], require at least three buffers.

15

In the following, the performance of the presented concepts by means of examples is discussed.

20

First, the processing is applied to an amplitude-panned mixture of 5 instrument recordings (drums, bass, keys, 2 guitars) sampled at 44100 Hz of which an excerpt of 3 seconds length is visualized. Drums, bass and keys are panned to the center, one guitar is panned to the left channel and the second guitar is panned to the right channel, both with $|ICLD| = 20\text{dB}$. A convolution reverb having stereo impulse responses with an RT60 of about 1.4 seconds per input channel is used to generate ambient signal components. The reverberated signal is added with a direct-to-ambient ratio of about 8 dB after K-weighting [29].

25

Fig. 6a-e show spectrograms the direct source signals and the left and right channel signals of the mixture signal. The spectrograms are computed using an STFT with a length of 2048 samples, 50 % overlap, a frame size of 1024 samples and a sine window. Please note that for the sake of clarity only the magnitudes of the spectral coefficients corresponding to frequencies up to 4 kHz are displayed. In particular, Fig. 6a-e illustrate input signals for the music example.

30

35

In particular, Fig. 6a-e illustrate in Fig. 6a source signals, wherein drums, bass and keys are panned to the center; in Fig. 6b source signals, wherein guitar 1, in the mix is panned to left; in Fig. 6c source signals wherein guitar 2, in the mix is panned to right; in Fig. 6d a left channel of a mixture signal; and in Fig. 6e a right channel of a mixture signal.

5

Fig. 7 shows the input signal and the output signal for the center signal extraction obtained by applying $G_{c2}(m, k; 1, 3)$. In particular, Fig. 7 is an example for center extraction, wherein input time signals (black) and output time signals (overlaid in gray) are illustrated, wherein Fig. 7, upper plot illustrates a left channel, and wherein Fig. 7, lower

10

The time constant for the recursive averaging in the PSD estimation here and in the following is set to 200 ms.

15

Fig. 8 illustrates the spectrograms of the output signal. Visual inspection reveals that the source signals panned off-center (shown in Fig. 6b and 6c) are largely attenuated in the output spectrograms. In particular, Fig. 8 illustrates an example for center extraction, more particularly spectrograms of the output signals. The output spectrograms also show that the ambient signal components are attenuated.

20

Fig. 9 shows the input signal and the output signal for the center signal attenuation obtained by applying $G_{s2}(m, k; 1, 3)$. The time signals illustrate that the transient sounds from the drums are attenuated by the processing. In particular, Fig. 9 illustrates an example for center attenuation, wherein input time signals (black) and output time signals

25

Fig. 10 illustrates the spectrograms of the output signal. It can be observed that the signals panned to the center are attenuated, for example when looking at the transient sound components and the sustained tones in the lower frequency range below 600Hz and comparing to Fig. 6a. The prominent sounds in the output signal correspond to the off-center panned instruments and the reverberation. In particular, Fig. 10 illustrates an example for center attenuation, more particularly, spectrograms of the output signals.

30

Informal listening over headphones reveals that the attenuation of the signal components is effective. When listening to the extracted center signal, processing artifacts become audible as slight modulations during the notes of guitar 2, similar to pumping in dynamic range compression. It can be noted that the reverberation is reduced and that the attenuation is more effective for low frequencies than for high frequencies. Whether this is

35

caused by the larger direct-to-ambient ratio in the lower frequencies, the frequency content of the sound sources or subjective perception due to unmasking phenomena can not be answered without a more detailed analysis.

- 5 When listening to the output signal where the center is attenuated, the overall sound quality is slightly better when compared to the center extraction result. Processing artifacts are audible as slight movements of the panned sources towards the center when dominant centered sources are active, equivalently to the pumping when extracting the center. The output signal sounds less direct as the result of the increased amount of
10 ambience in the output signal.

To illustrate the PDC filtering, Fig. 11a-d show two speech signals which have been mixed to obtain input signals with and without ICTD. In particular, Fig. 11a-d illustrate input source signals for illustrating the PDC, wherein Fig. 11a illustrates source signal 1;
15 wherein Fig. 11b illustrates source signal 2; wherein Fig. 11c illustrates a left channel of a mixture signal; and wherein Fig. 11d illustrates a right channel of a mixture signal.

The two-channel mixture signal is generated by mixing the speech source signals with equal gains to each channel and by adding white noise with an SNR of 10 dB (K-
20 weighted) to the signal.

Fig. 12a-c show the spectral weights computed from gain function (13). In particular, Fig. 12a-c illustrate spectral weights $G_{c2}(m, k, 1, 3)$ for demonstrating the PDC filtering, wherein Fig. 12a illustrates spectral weights for input signals without ICTD, PDC disabled;
25 Fig. 12b illustrates spectral weights for input signals with ICTD, PDC disabled; and Fig. 12c illustrates spectral weights for input signals with ICTD, PDC enabled.

The spectral weights in the upper plot are close to 0 dB when speech is active and assume the minimum value in time-frequency regions with low SNR. The second plot
30 shows the spectral weights for an input signal where the first speech signal (Fig. 11a) is mixed with an ICTD of 26 samples. The comb-filter characteristics is illustrated in Fig. 12b. Fig. 12c shows the spectral weights when PDC is enabled. The comb-filtering artifacts are largely reduced, although the compensation is not perfect near the notch frequencies at 848Hz and 2544Hz.

35

Informal listening shows that the additive noise is largely attenuated. When processing signals without ICTD, the output signals have a bit of an ambient sound characteristic which results presumably from the phase incoherence introduced by the additive noise.

When processing signals with ICTD, the first speech signal (Fig. 11a) is largely attenuated and strong comb-filtering artifacts are audible when not applying the PDC filtering. With additional PDC filtering, the comb-filtering artifacts are still slightly audible, but much less annoying. Informal listening to other material reveals light artifacts, which can be reduced either by decreasing γ , by increasing β , or by adding a scaled version of the unprocessed input signal to the output. In general, artifacts are less audible when attenuating the center signal and more audible when extracting the center signal. Distortions of the perceived spatial image are very small. This can be attributed to the fact that the spectral weights are identical for all channel signals and do not affect the ICLDs. The comb-filtering artifacts are hardly audible when processing natural recordings featuring time-of-arrival stereophony for whom a mono downmix is not subject to strong audible comb-filtering artifacts. For the PDC filtering it can be noted that small values of the time constant of the recursive averaging (in particular the instantaneous compensation of phase differences when computing X_d) introduces coherence in the signals used for the downmix. Consequently, the processing is agnostic with respect to the diffuseness of the input signal. When the time constant is increased, it can be observed that (1) the effect of the PDC for input signals with amplitude difference stereophony decreases and (2) the comb-filtering effect becomes more audible at note onsets when the direct sound sources are not time-aligned between the input channels.

Concepts for scaling the center signal in audio recordings by applying real-valued spectral weights which are computed from monotonic functions of the SDR have been provided. The rationale is that center signal scaling needs to take into account both, the lateral displacement of direct sources and the amount of diffuseness, and that these characteristics are implicitly captured by the SDR. The processing can be controlled by semantically meaningful user parameters and is in comparison to other frequency domain techniques of low computational complexity and memory load. The proposed concepts give good results when processing input signals featuring amplitude difference stereophony, but can be subject to comb-filtering artifacts when the direct sound sources are not time-aligned between the input channels. A first approach to solve this is to compensate for non-zero phase in the inter-channel transfer function.

So far, the concepts of embodiments have been tested by means of informal listening. For typical commercial recordings, the results are of good sound quality but also depend on the desired separation strength.

Although some aspects have been described in the context of an apparatus, it is clear that these aspects also represent a description of the corresponding method, where a block or

device corresponds to a method step or a feature of a method step. Analogously, aspects described in the context of a method step also represent a description of a corresponding block or item or feature of a corresponding apparatus.

- 5 The inventive decomposed signal can be stored on a digital storage medium or can be transmitted on a transmission medium such as a wireless transmission medium or a wired transmission medium such as the Internet.

10 Depending on certain implementation requirements, embodiments of the invention can be implemented in hardware or in software. The implementation can be performed using a digital storage medium, for example a floppy disk, a DVD, a CD, a ROM, a PROM, an EPROM, an EEPROM or a FLASH memory, having electronically readable control signals stored thereon, which cooperate (or are capable of cooperating) with a programmable computer system such that the respective method is performed.

15 Some embodiments according to the invention comprise a non-transitory data carrier having electronically readable control signals, which are capable of cooperating with a programmable computer system, such that one of the methods described herein is performed.

20 Generally, embodiments of the present invention can be implemented as a computer program product with a program code, the program code being operative for performing one of the methods when the computer program product runs on a computer. The program code may for example be stored on a machine readable carrier.

25 Other embodiments comprise the computer program for performing one of the methods described herein, stored on a machine readable carrier.

30 In other words, an embodiment of the inventive method is, therefore, a computer program having a program code for performing one of the methods described herein, when the computer program runs on a computer.

35 A further embodiment of the inventive methods is, therefore, a data carrier (or a digital storage medium, or a computer-readable medium) comprising, recorded thereon, the computer program for performing one of the methods described herein.

A further embodiment of the inventive method is, therefore, a data stream or a sequence of signals representing the computer program for performing one of the methods

described herein. The data stream or the sequence of signals may for example be configured to be transferred via a data communication connection, for example via the Internet.

- 5 A further embodiment comprises a processing means, for example a computer, or a programmable logic device, configured to or adapted to perform one of the methods described herein.

- 10 A further embodiment comprises a computer having installed thereon the computer program for performing one of the methods described herein.

- In some embodiments, a programmable logic device (for example a field programmable gate array) may be used to perform some or all of the functionalities of the methods described herein. In some embodiments, a field programmable gate array may cooperate
15 with a microprocessor in order to perform one of the methods described herein. Generally, the methods are preferably performed by any hardware apparatus.

- The above described embodiments are merely illustrative for the principles of the present invention. It is understood that modifications and variations of the arrangements and the
20 details described herein will be apparent to others skilled in the art. It is the intent, therefore, to be limited only by the scope of the impending patent claims and not by the specific details presented by way of description and explanation of the embodiments herein.

References:

- [1] International Telecommunication Union, Radiocommunication Assembly, "Multichannel stereophonic sound system with and without accompanying picture.," Recommendation ITU-R BS.775-2, 2006, Geneva, Switzerland.
- [2] J. Berg and F. Rumsey, "Identification of quality attributes of spatial sound by repertory grid technique," *J. Audio Eng. Soc.*, vol. 54, pp. 365–379, 2006.
- [3] J. Blauert, *Spatial Hearing*, MIT Press, 1996.
- [4] F. Rumsey, "Controlled subjective assessment of two-to-five channel surround sound processing algorithms," *J. Audio Eng. Soc.*, vol. 47, pp. 563–582, 1999.
- [5] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue enhancement - technology and experiments," *EBU Technical Review*, vol. Q2, pp. 1–11, 2012.
- [6] J.-H. Bach, J. Anemüller, and B. Kollmeier, "Robust speech detection in real acoustic backgrounds with perceptually motivated features," *Speech Communication*, vol. 53, pp. 690–706, 2011.
- [7] C. Avendano and J.-M. Jot, "A frequency-domain approach to multi-channel upmix," *J. Audio Eng. Soc.*, vol. 52, 2004.
- [8] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2004.
- [9] E. Vickers, "Two-to-three channel upmix for center channel derivation and speech enhancement," in *Proc. Audio Eng. Soc. 127th Conv.*, 2009.
- [10] D. Jang, J. Hong, H. Jung, and K. Kang, "Center channel separation based on spatial analysis," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2008.
- [11] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2000.

- [12] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Proc.*, vol. 52, pp. 1830–1847, 2004.
- 5 [13] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer, 2007.
- [14] N. Cahill, R. Cooney, K. Humphreys, and R. Lawlor, "Speech source enhancement using a modified ADress algorithm for applications in mobile communications," in *Proc. Audio Eng. Soc. 121st Conv.*, 2006.
- 10 [15] M. Puigt and Y. Deville, "A time-frequency correlation-based blind source separation method for time-delay mixtures," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006.
- 15 [16] Simon Arberet, Remi Gribonval, and Frederic Bimbot, "A robust method to count and locate audio sources in a stereophonic linear anechoic mixture," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007.
- 20 [17] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 18, pp. 382–394, 2010.
- [18] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2003.
- 25 [19] A. Favrot, M. Erne, and C. Faller, "Improved cocktail-party processing," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2006.
- [20] US patent 7,630,500 B1, P.E. Beckmann, 2009
- 30 [21] US patent 7,894,611 B2, P.E. Beckmann, 2011
- [22] J.B. Allen, D.A. Berkeley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, 1977.
- 35 [23] J. Merimaa, M. Goodwin, and J.-M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. Audio Eng. Soc. 123rd Conv.*, 2007.

- 5
- [24] J. Usher and J. Benesty, "Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 2141–2150, 2007.
- [25] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, vol. 54, 2006.
- 10 [26] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, "Ambience separation from mono recordings using Non-negative Matrix Factorization," in *Proc. Audio Eng. Soc. 30th Int. Conf.*, 2007.
- 15 [27] C. Uhle and C. Paul, "A supervised learning approach to ambience extraction from mono recordings for blind upmixing," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2008.
- [28] G. Soulodre, "System for extracting and changing the reverberant content of an audio input signal," US Patent 8,036,767, Oct. 2011.
- 20 [29] International Telecommunication Union, Radiocommunication Assembly, "Algorithms to measure audio programme loudness and true-peak audio level," Recommendation ITUR BS.1770-2, March 2011, Geneva, Switzerland.

Claims

1. An apparatus for generating a modified audio signal comprising two or more modified audio channels from an audio input signal comprising two or more audio input channels, wherein the apparatus comprises:

an information generator for generating signal-to-downmix information, wherein the information generator is adapted to generate signal information by combining a spectral value of each of the two or more audio input channels, wherein the information generator is adapted to generate downmix information by combining the spectral value of each of the two or more audio input channels, and wherein the information generator is adapted to combine the signal information and the downmix information to obtain signal-to-downmix information, and

a signal attenuator for attenuating the two or more audio input channels depending on the signal-to-downmix information to obtain the two or more modified audio channels,

wherein the information generator is configured to generate the signal information $\Phi_1(m, k)$ according to the formula

$$\Phi_1(m, k) = \mathcal{E}\{\mathbf{W}\mathbf{X}(m, k)(\mathbf{W}\mathbf{X}(m, k))^H\}$$

wherein the information generator is configured to generate the downmix information $\Phi_2(m, k)$ according to the formula

$$\Phi_2(m, k) = \mathcal{E}\{\mathbf{V}\mathbf{X}(m, k)(\mathbf{V}\mathbf{X}(m, k))^H\}, \text{ and}$$

wherein the information generator is configured to generate a signal-to-downmix ratio as the signal-to-downmix information $R_g(m, k, \beta)$ according to the formula

$$R_g(m, k, \beta) = \left(\frac{\text{tr}\{\Phi_1(m, k)^\beta\}}{\text{tr}\{\Phi_2(m, k)^\beta\}} \right)^{\frac{1}{2\beta-1}}$$

wherein $\mathbf{X}(m, k)$ indicates the audio input signal, wherein

$$\mathbf{X}(m, k) = [X_1(m, k) \cdots X_N(m, k)]^T,$$

wherein N indicates the number of audio input channels of the audio input signal,

5 wherein m indicates a time index, and wherein k indicates a frequency index,

wherein $X_1(m, k)$ indicates the first audio input channel, wherein $X_N(m, k)$ indicates the N -th audio input channel,

10 wherein $\mathbf{V}$ indicates a matrix or a vector,

wherein $\mathbf{W}$ indicates a matrix or a vector,

wherein ^{H} indicates the conjugate transpose of a matrix or a vector,

15 wherein $\mathcal{E}\{\cdot\}$ is an expectation operation,

wherein β is a real number with $\beta > 0$, and

20 wherein $\text{tr}\{\cdot\}$ is the trace of a matrix.

2. An apparatus according to claim 1, wherein $\mathbf{V}$ is a row vector of length N whose elements are equal to one and $\mathbf{W}$ is the identity matrix of size $N \times N$.

25 3. An apparatus according to any one of claims 1 or 2, wherein $\mathbf{V} = [1, 1]$, wherein $\mathbf{W} = [1, -1]$ and wherein $N = 2$.

4. An apparatus according to any one of claims 1 to 3, wherein the number of the modified audio channels is equal to the number of the audio input channels, or
30 wherein the number of the modified audio channels is smaller than the number of the audio input channels.

5. An apparatus according to any one of claims 1 to 4,

35 wherein the information generator is configured to process the spectral value of each of the two or more audio input channels to obtain two or more processed

values, and wherein the information generator is configured to combine the two or more processed values to obtain the signal information, and

5 wherein the information generator is configured to combine the spectral value of each of the two or more audio input channels to obtain a combined value, and wherein the information generator is configured to process the combined value to obtain the downmix information.

6. 10 An apparatus according to any one of claims 1 to 5, wherein the information generator is configured to process the spectral value of each of the two or more audio input channels by multiplying said spectral value by the complex conjugate of said spectral value to obtain an auto power spectral density of said spectral value for each of the two or more audio input channels.

15 7. An apparatus according to claim 5,

wherein the information generator is configured to process the spectral value of each of the two or more audio input channels by multiplying said spectral value by the complex conjugate of said spectral value to obtain an auto power spectral density of said spectral value for each of the two or more audio input channels,

20

wherein the information generator is configured to process the combined value by determining a power spectral density of the combined value.

25 8. An apparatus according to claim 7,

wherein the information generator is configured to determine the signal information

$$\Phi_1(m, k) = \mathcal{E}\{\mathbf{W}\mathbf{X}(m, k)(\mathbf{W}\mathbf{X}(m, k))^H\}$$

30

according to the formula

$$s(m, k, \beta) = \sum_{i=1}^N \Phi_{i,i}(m, k)^\beta,$$

35

wherein $\Phi_{i,i}(m, k)$ indicates the auto power spectral density of the spectral value of an i -th audio signal channel.

9. An apparatus according to claim 8,

5

wherein the information generator is configured to determine the signal-to-downmix ratio

$$R_g(m, k, \beta) = \left(\frac{\text{tr}\{\Phi_1(m, k)^\beta\}}{\text{tr}\{\Phi_2(m, k)^\beta\}} \right)^{\frac{1}{2\beta-1}}$$

10

according to the formula

$$R(m, k, \beta) = \left(\frac{\sum_{i=1}^N \Phi_{i,i}(m, k)^\beta}{\Phi_d(m, k)^\beta} \right)^{\frac{1}{2\beta-1}}$$

- 15 to obtain the signal-to-downmix ratio,

wherein $\Phi_d(m, k)$ indicates the power spectral density of the combined value.

10. An apparatus according to any one of claims 1 to 9, wherein the signal attenuator is adapted to attenuate the two or more audio input channels depending on a gain function $G(m, k)$ according to the formula

20

$$Y(m, k) = G(m, k)X(m, k),$$

- 25 wherein the gain function $G(m, k)$ depends on the signal-to-downmix information, and wherein the gain function $G(m, k)$ is a monotonically increasing function of the signal-to-downmix information or a monotonically decreasing function of the signal-to-downmix information,

- 30 wherein $X(m, k)$ indicates the audio input signal,

wherein $Y(m, k)$ indicates the modified audio signal,

wherein m indicates a time index, and

wherein k indicates a frequency index.

11. An apparatus according to claim 10,

5 wherein the gain function $G(m, k)$ is a first function $G_{c1}(m, k, \beta, \gamma)$, a second function $G_{c2}(m, k, \beta, \gamma)$, a third function $G_{s1}(m, k, \beta, \gamma)$ or a fourth function $G_{s2}(m, k, \beta, \gamma)$,

wherein

10

$$G_{c1}(m, k, \beta, \gamma) = (1 + R_{\min} - R(m, k, \beta))^{\gamma},$$

wherein

15

$$G_{c2}(m, k, \beta, \gamma) = \left(\frac{R_{\min}}{R(m, k, \beta)} \right)^{\gamma},$$

wherein

20

$$G_{s1}(m, k, \beta, \gamma) = R(m, k, \beta)^{\gamma},$$

wherein

$$G_{s2}(m, k, \beta, \gamma) = \left(1 + R_{\min} - \frac{R_{\min}}{R(m, k, \beta)} \right)^{\gamma},$$

25

wherein β is a real number with $\beta > 0$,

wherein γ is a real number with $\gamma > 0$, and

wherein $R_{\min}$ indicates the minimum of R .

30

12. A system comprising:

35

a phase compensator for generating a phase-compensated audio signal comprising two or more phase-compensated audio channels from an unprocessed audio signal comprising two or more unprocessed audio channels, and

an apparatus according to any one of claims 1 to 11 adapted to: receive, as the audio input signal, the phase-compensated audio signal generated by the phase compensator; and generate the modified audio signal comprising the two or more modified audio channels from the audio input signal comprising the two or more phase-compensated audio channels as the two or more audio input channels,

wherein one of the two or more unprocessed audio channels is a reference channel,

wherein the phase compensator is adapted to estimate for each unprocessed audio channel of the two or more unprocessed audio channels which is not the reference channel a phase transfer function between said unprocessed audio channel and the reference channel, and

wherein the phase compensator is adapted to generate the phase-compensated audio signal by modifying each unprocessed audio channel of the unprocessed audio channels which is not the reference channel depending on the phase transfer function of said unprocessed audio channel.

13. A method for generating a modified audio signal comprising two or more modified audio channels from an audio input signal comprising two or more audio input channels, wherein the method comprises:

generating signal information by combining a spectral value of each of the two or more audio input channels,

generating downmix information by combining the spectral value of each of the two or more audio input channels,

generating signal-to-downmix information by combining the signal information and the downmix information, and

attenuating the two or more audio input channels depending on the signal-to-downmix information to obtain the two or more modified audio channels,

wherein generating the signal information $\Phi_1(m, k)$ is conducted according to the formula

$$\Phi_1(m, k) = \mathcal{E}\{\mathbf{W}\mathbf{X}(m, k)(\mathbf{W}\mathbf{X}(m, k))^H\}$$

wherein generating the downmix information $\Phi_2(m, k)$ is conducted according to the formula

$$\Phi_2(m, k) = \mathcal{E}\{\mathbf{V}\mathbf{X}(m, k)(\mathbf{V}\mathbf{X}(m, k))^H\}, \text{ and}$$

wherein a signal-to-downmix ratio is generated as the signal-to-downmix information $R_g(m, k, \beta)$ according to the formula

$$R_g(m, k, \beta) = \left(\frac{\text{tr}\{\Phi_1(m, k)^\beta\}}{\text{tr}\{\Phi_2(m, k)^\beta\}} \right)^{\frac{1}{2\beta-1}}$$

wherein $\mathbf{X}(m, k)$ indicates the audio input signal, wherein

$$\mathbf{X}(m, k) = [X_1(m, k) \cdots X_N(m, k)]^T,$$

wherein N indicates the number of audio input channels of the audio input signal,

wherein m indicates a time index, and wherein k indicates a frequency index,

wherein $X_1(m, k)$ indicates the first audio input channel, wherein $X_N(m, k)$ indicates the N -th audio input channel,

wherein $\mathbf{V}$ indicates a matrix or a vector,

wherein $\mathbf{W}$ indicates a matrix or a vector,

wherein ^{H} indicates the conjugate transpose of a matrix or a vector,

wherein $\mathcal{E}\{\cdot\}$ is an expectation operation,

wherein β is a real number with $\beta > 0$, and

wherein $\text{tr}\{\cdot\}$ is the trace of a matrix.

14. A computer program product comprising a computer readable memory storing computer executable instructions thereon that when executed by a computer perform the method of claim 13.

5

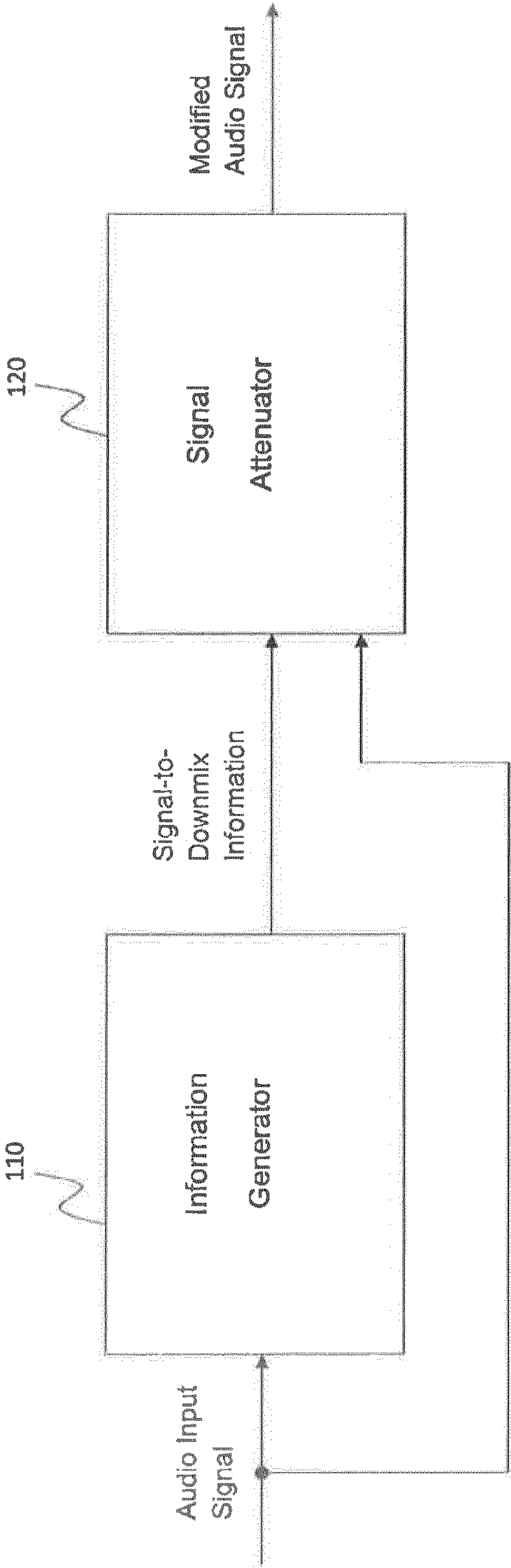


FIG 1

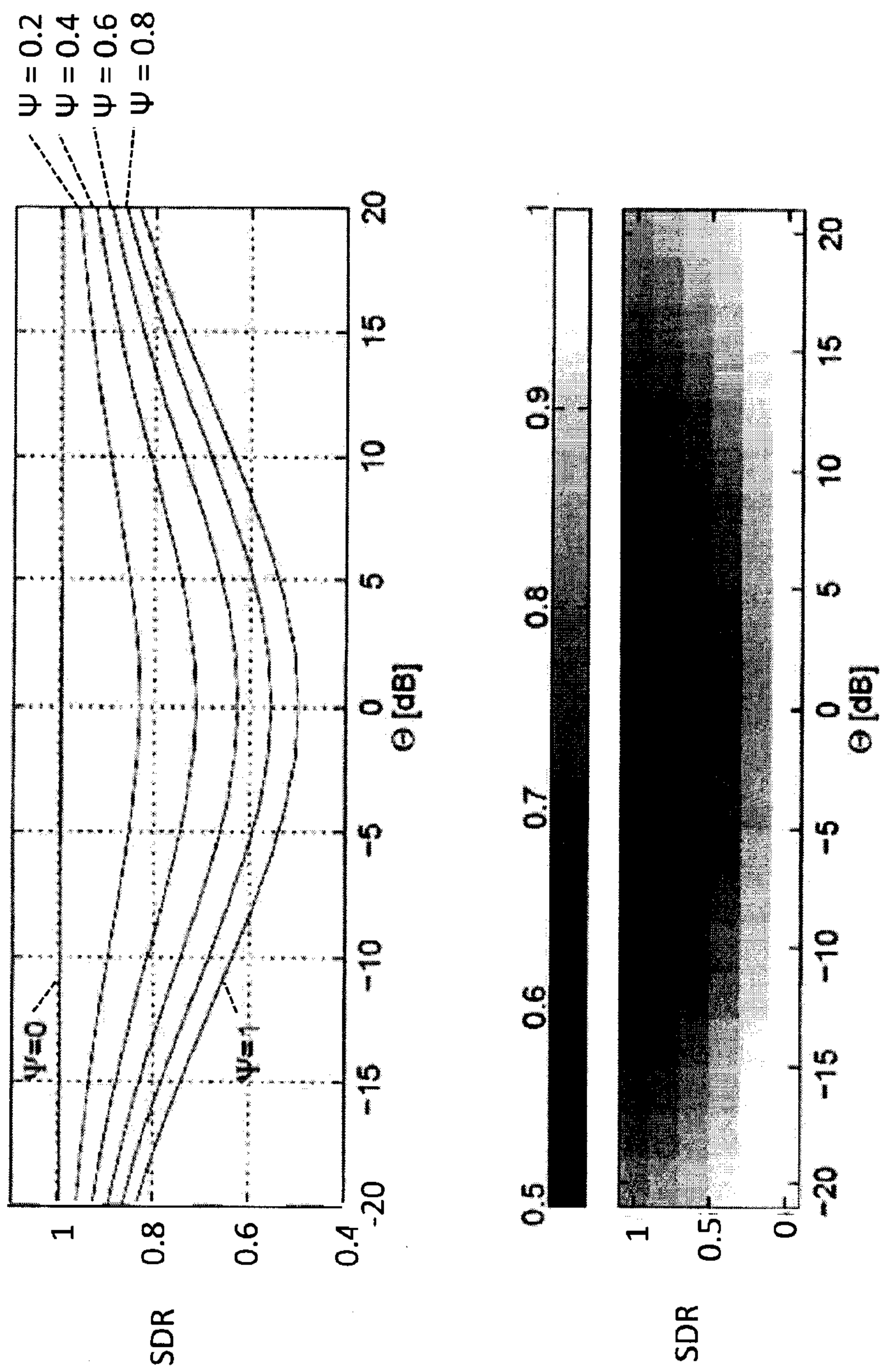


FIG 2

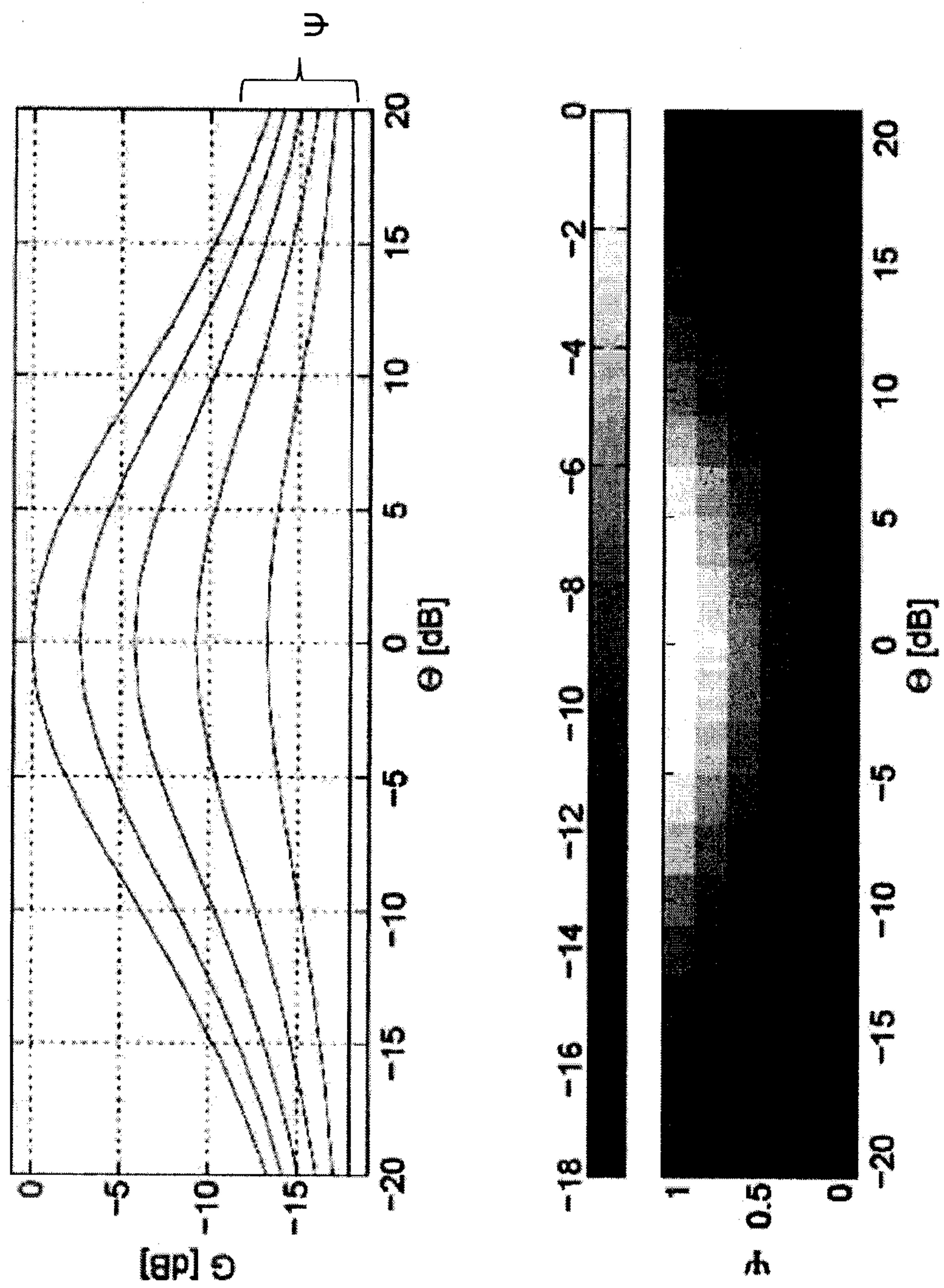


FIG 3

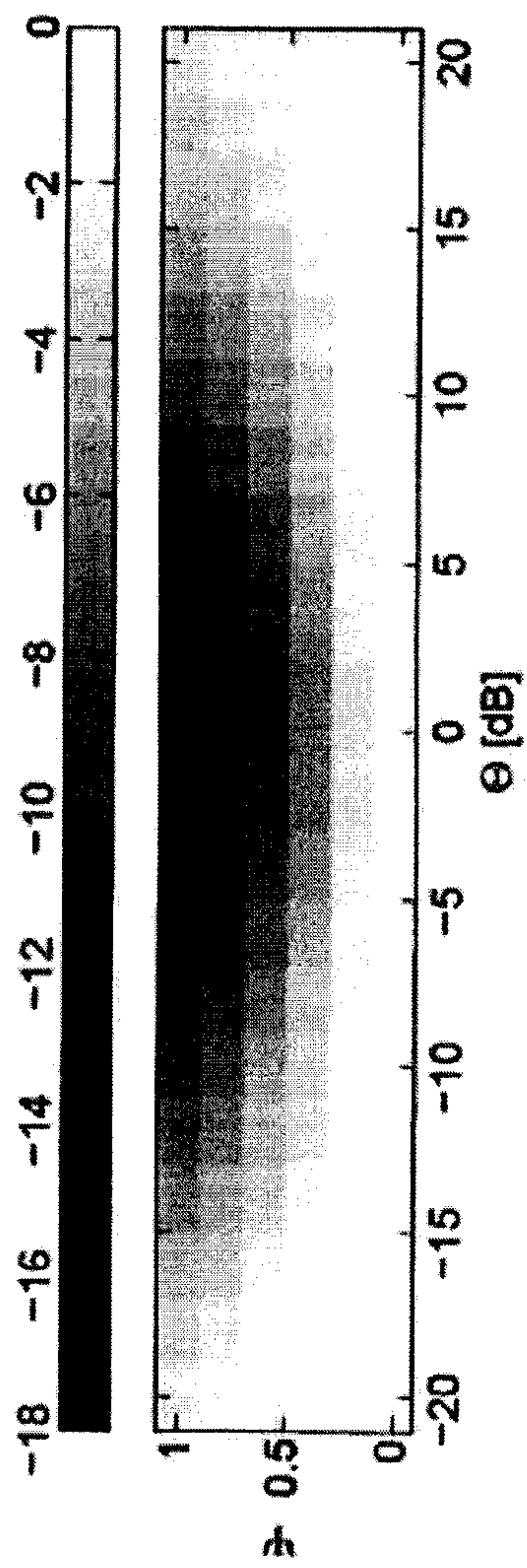
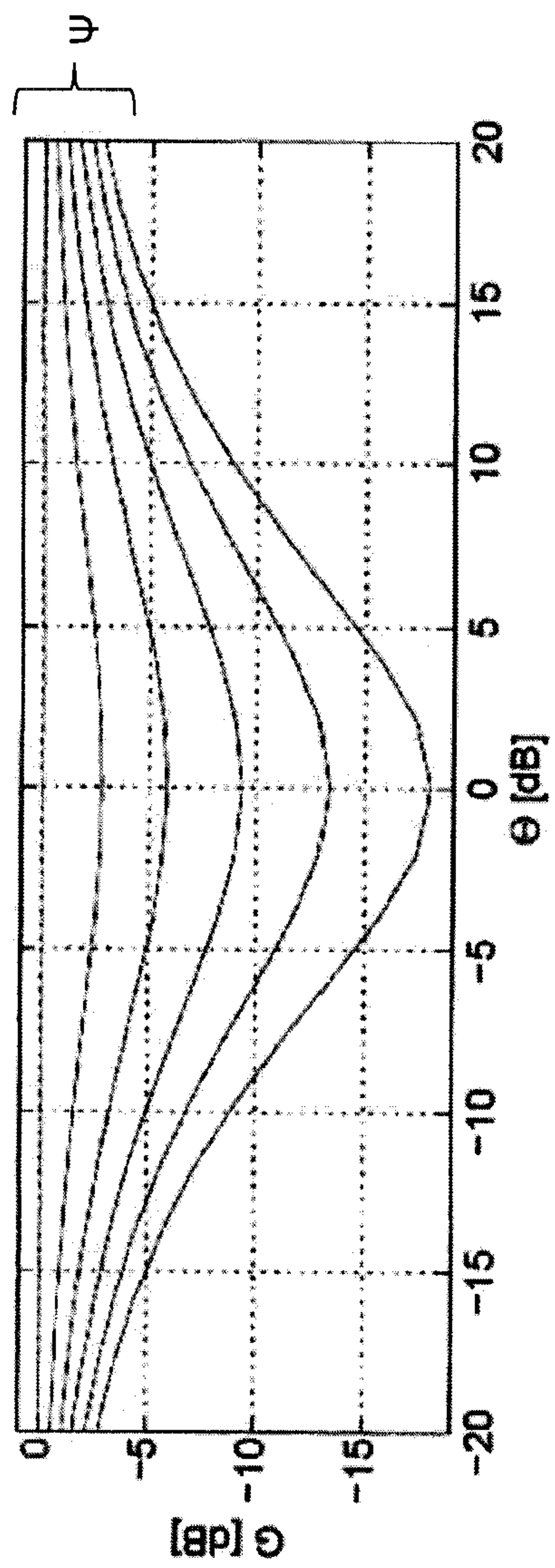


FIG 4

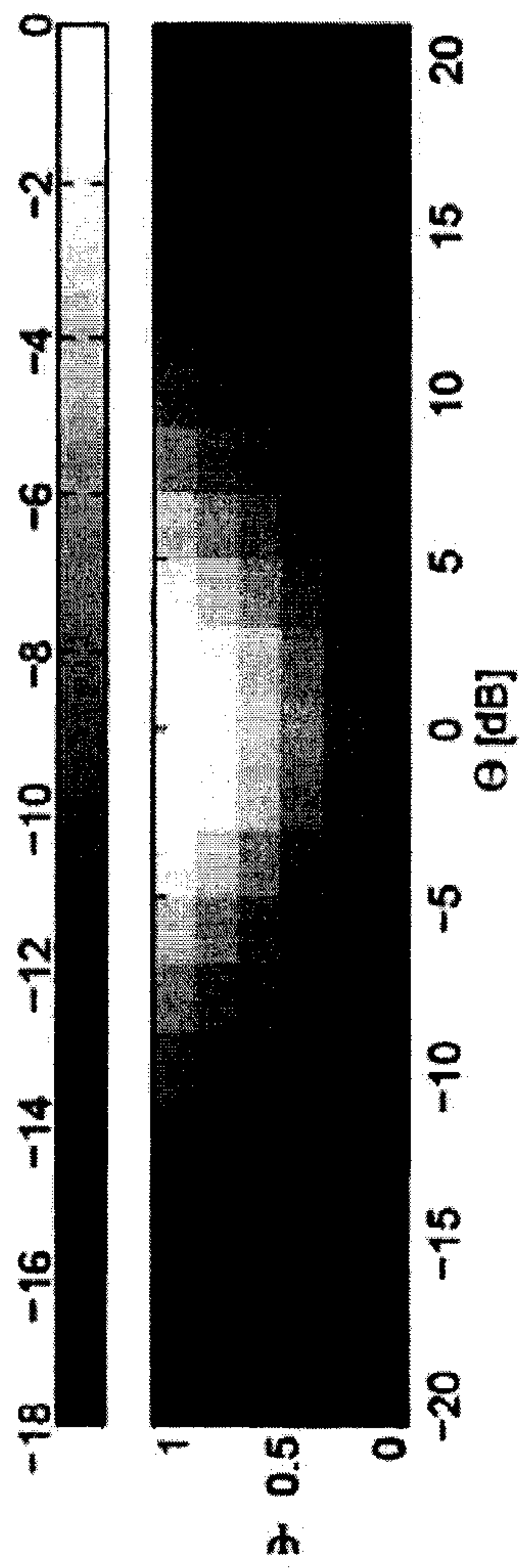
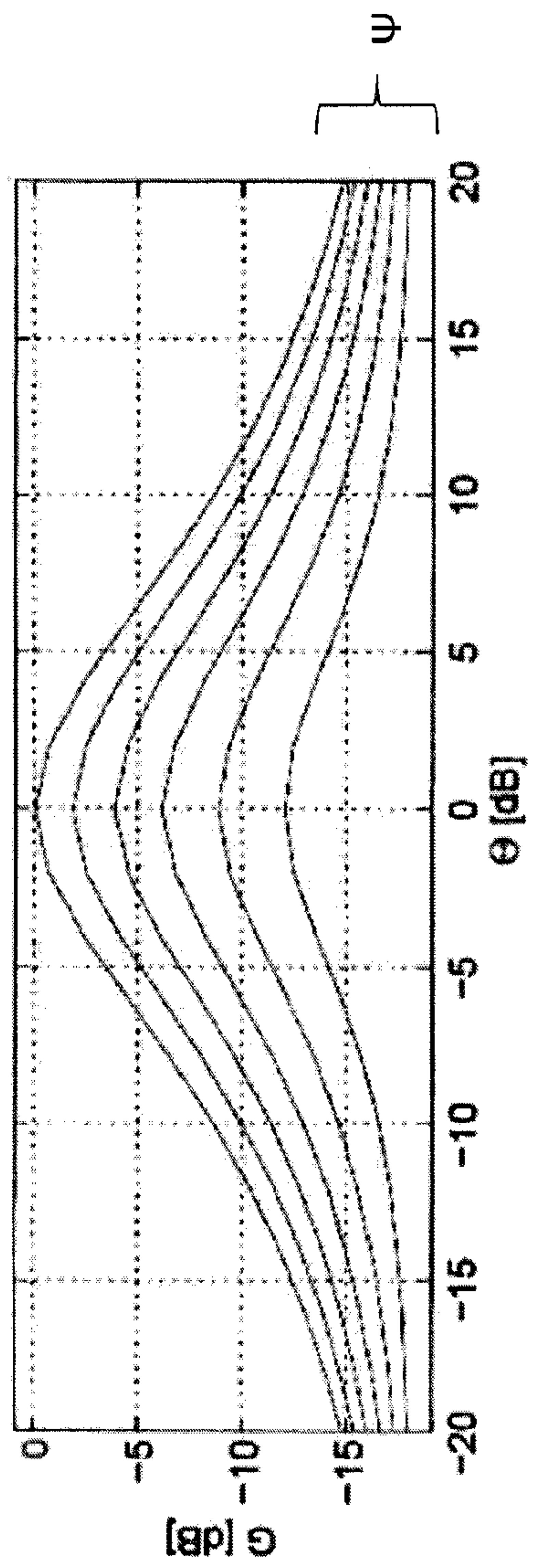


FIG 5

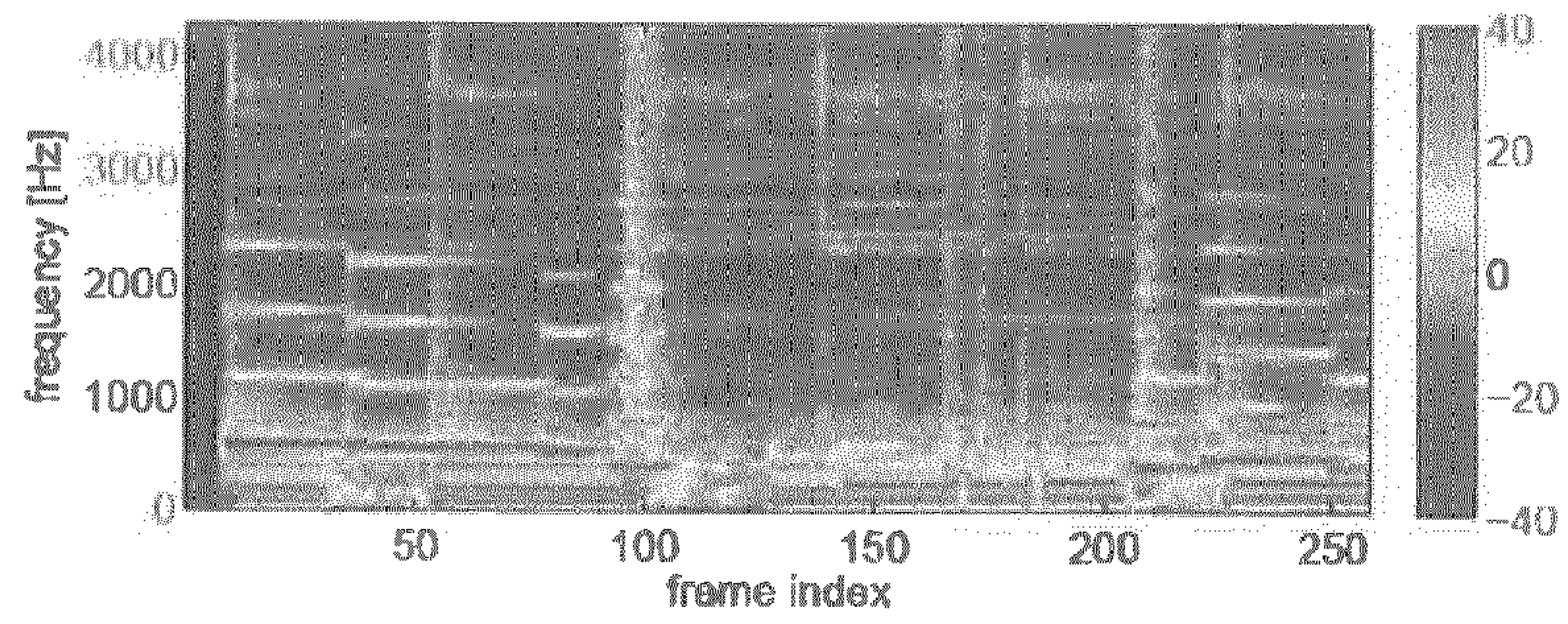


FIG 6A

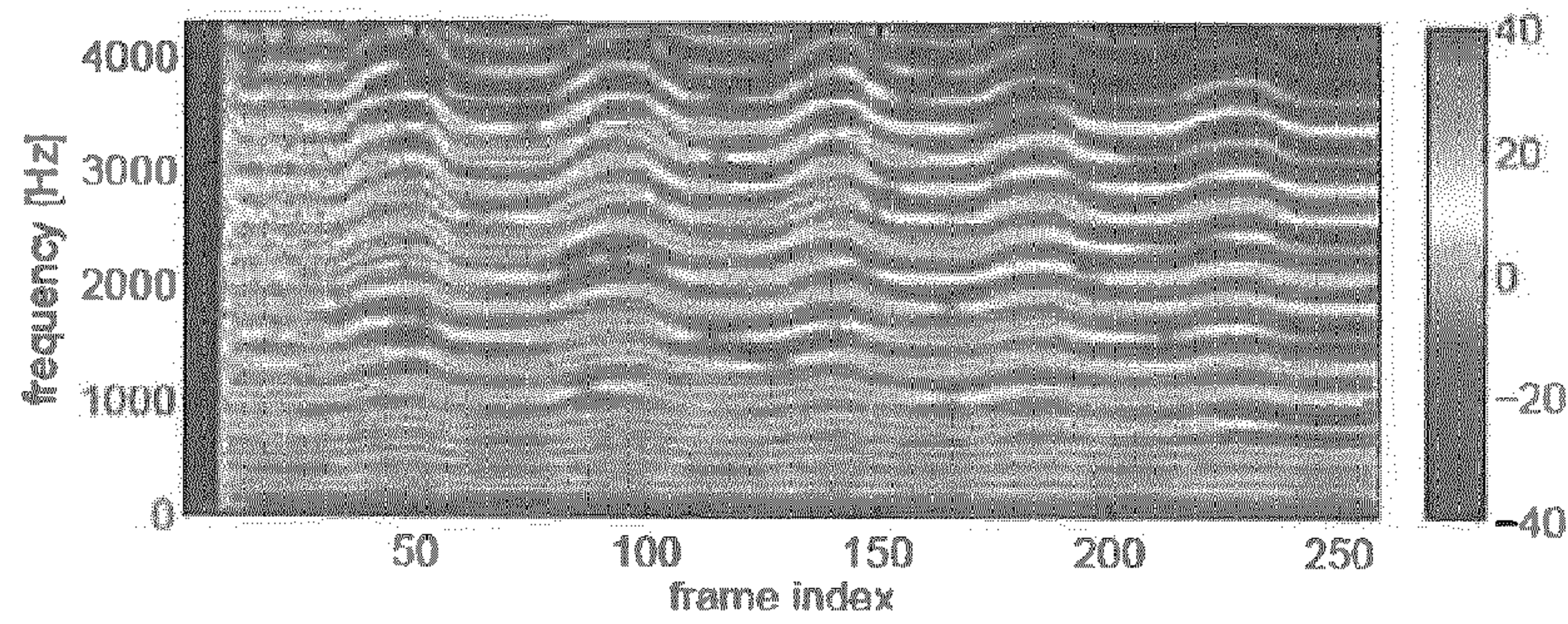


FIG 6B

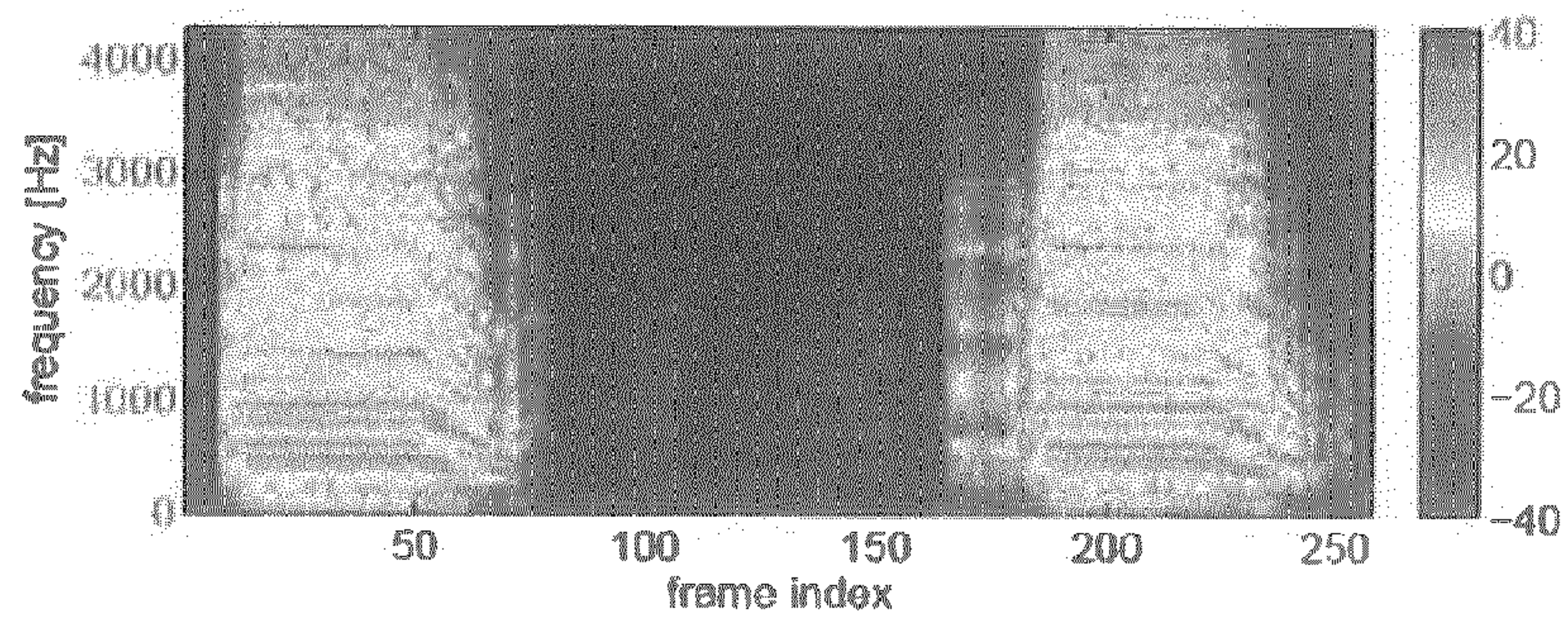


FIG 6C

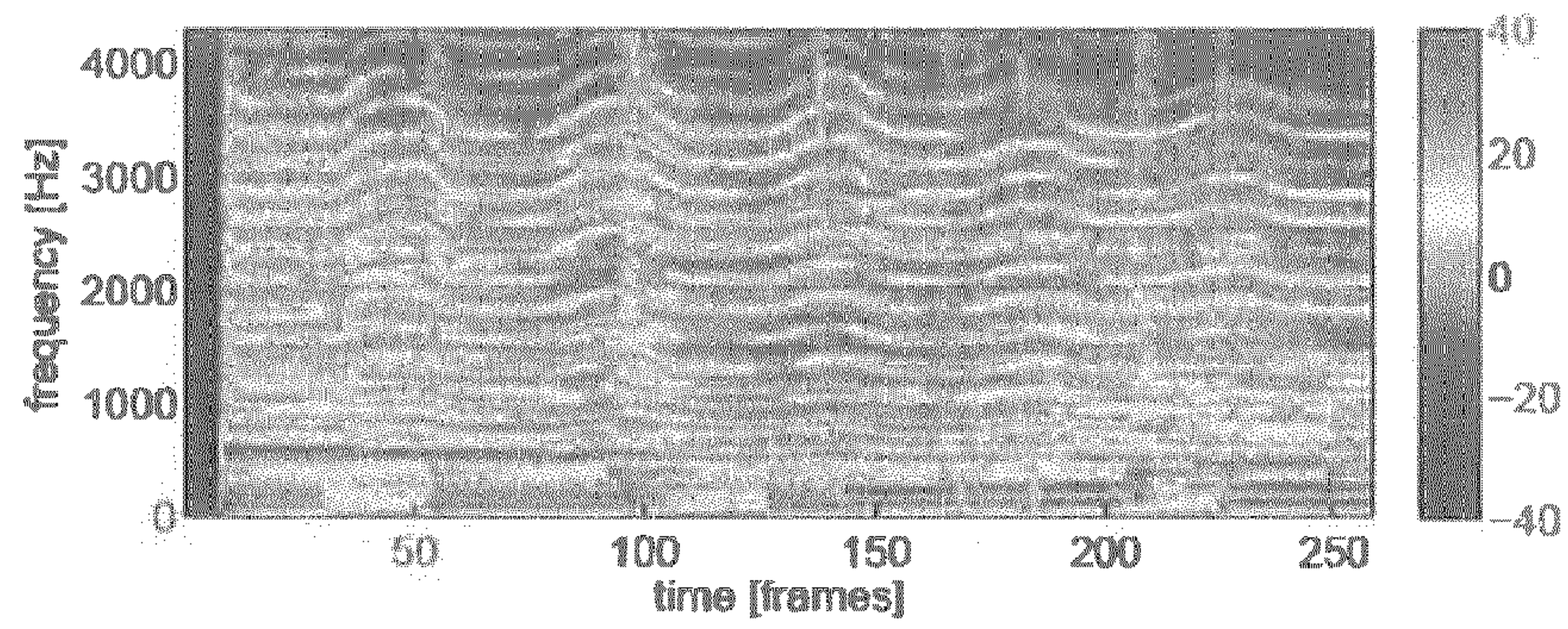


FIG 6D

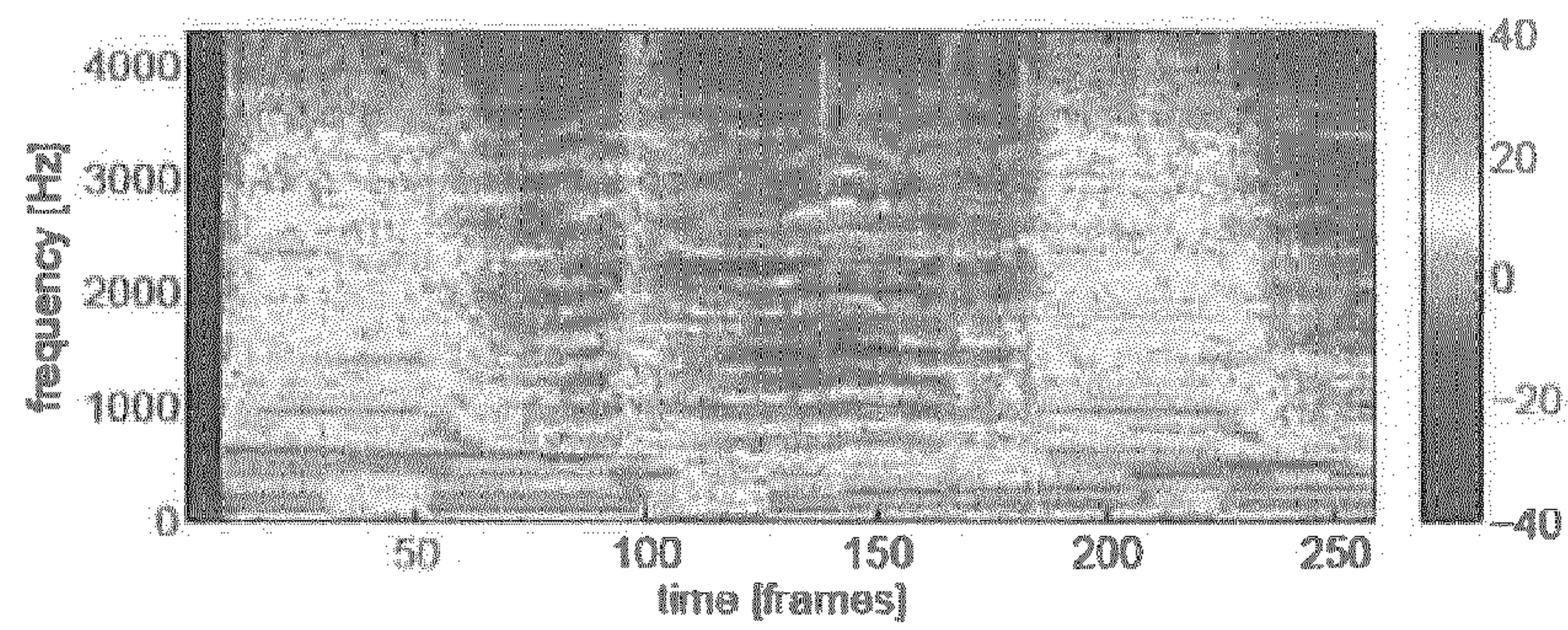


FIG 6E

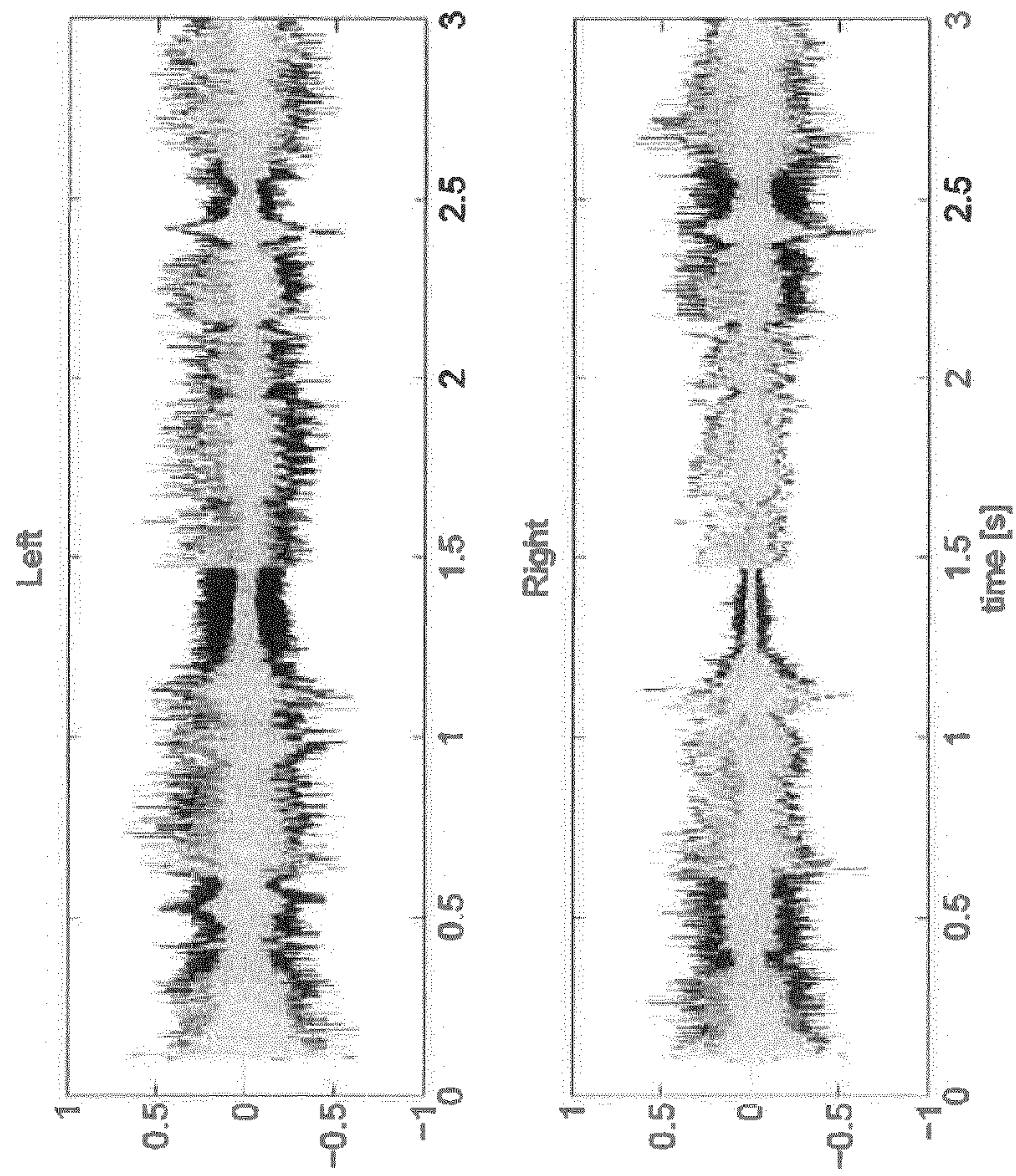


FIG 7

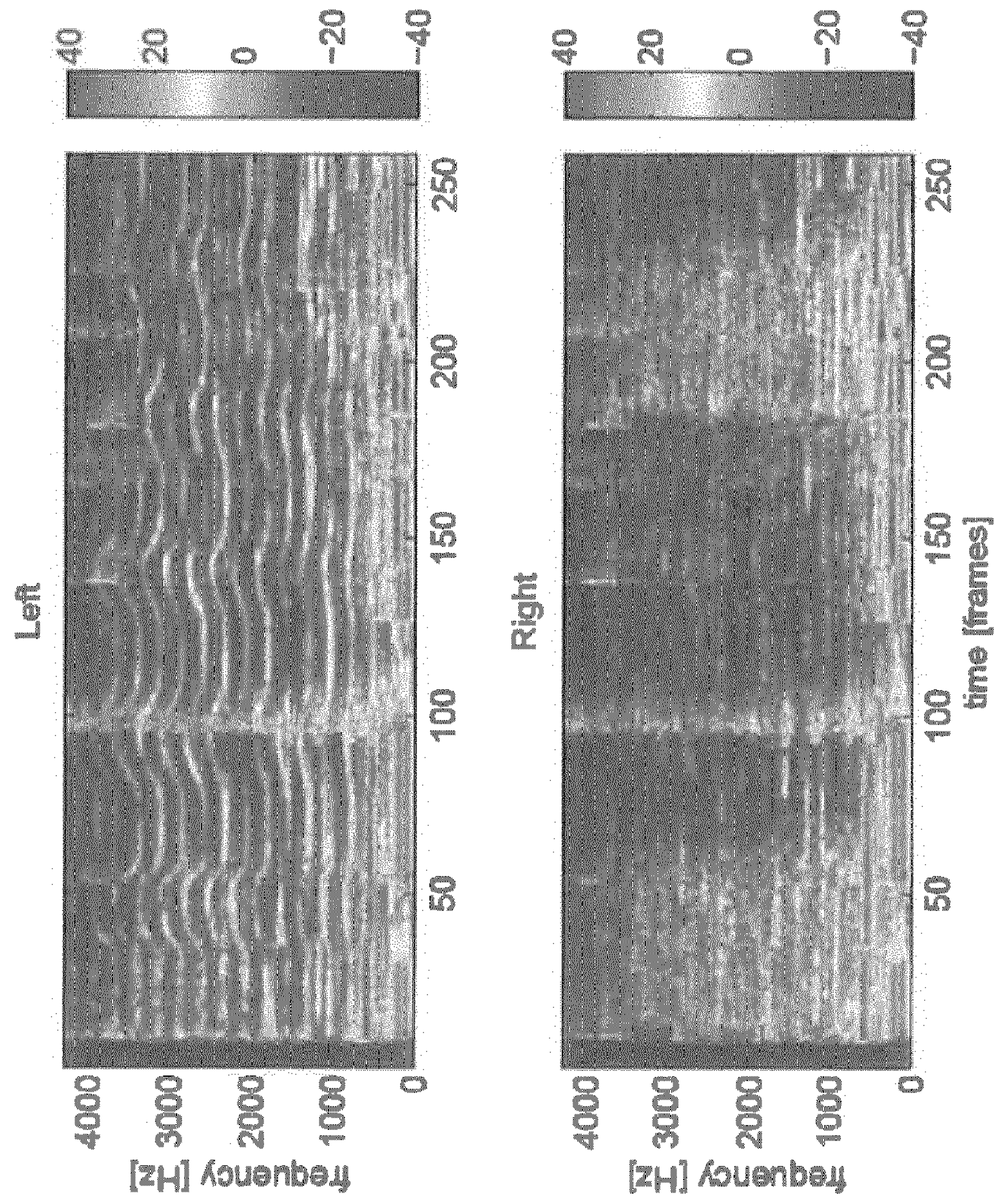


FIG 8

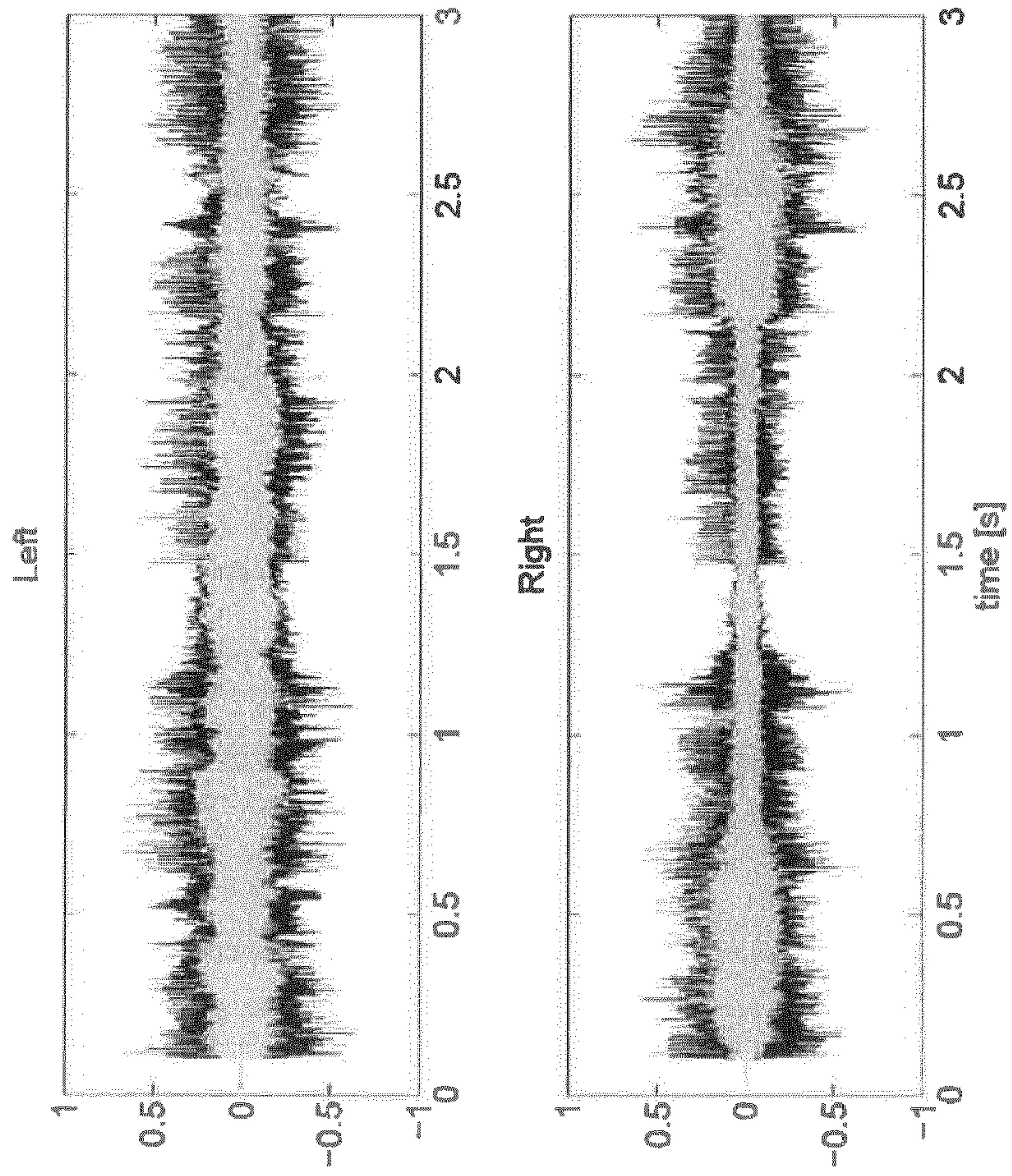


FIG 9

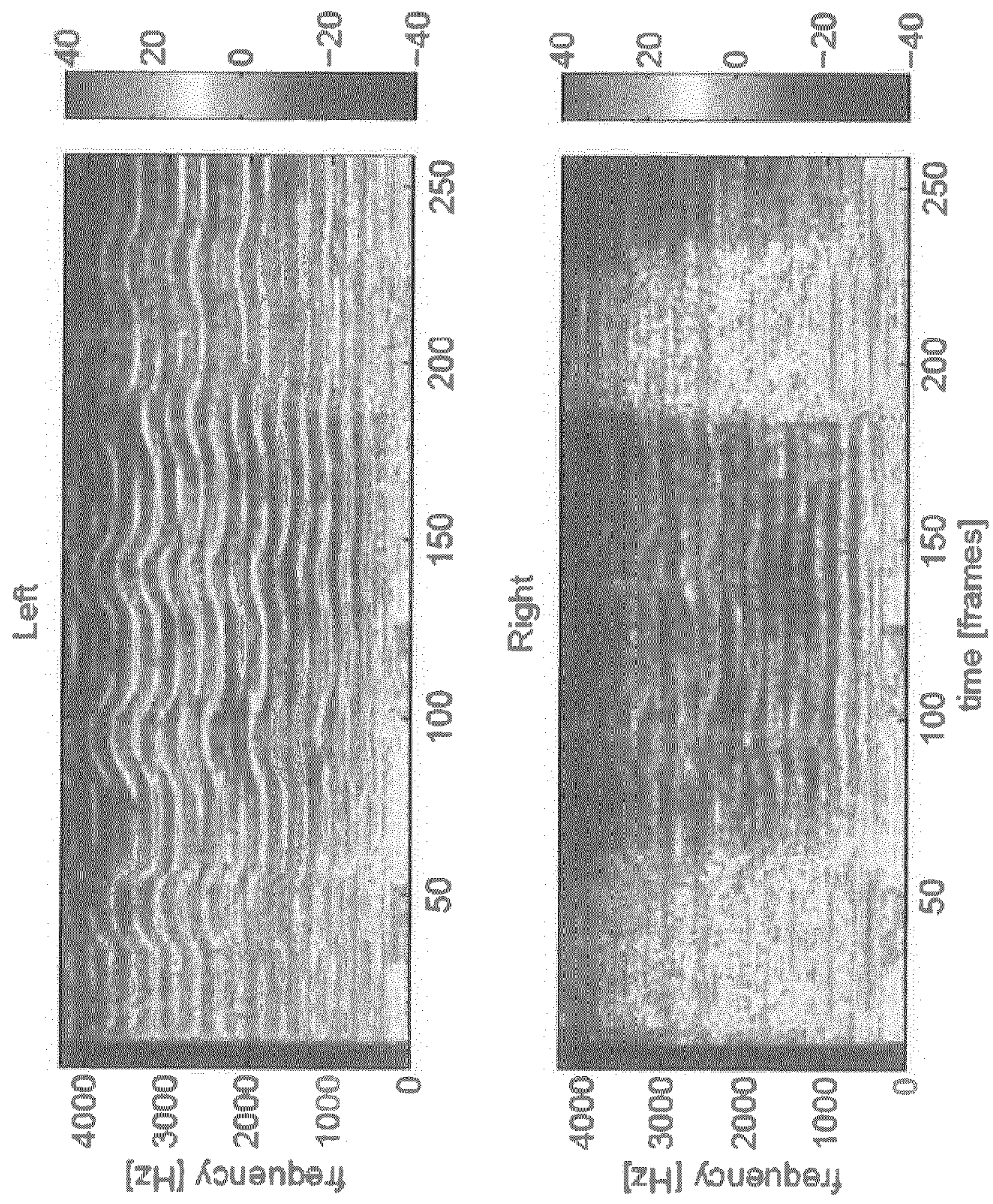


FIG 10

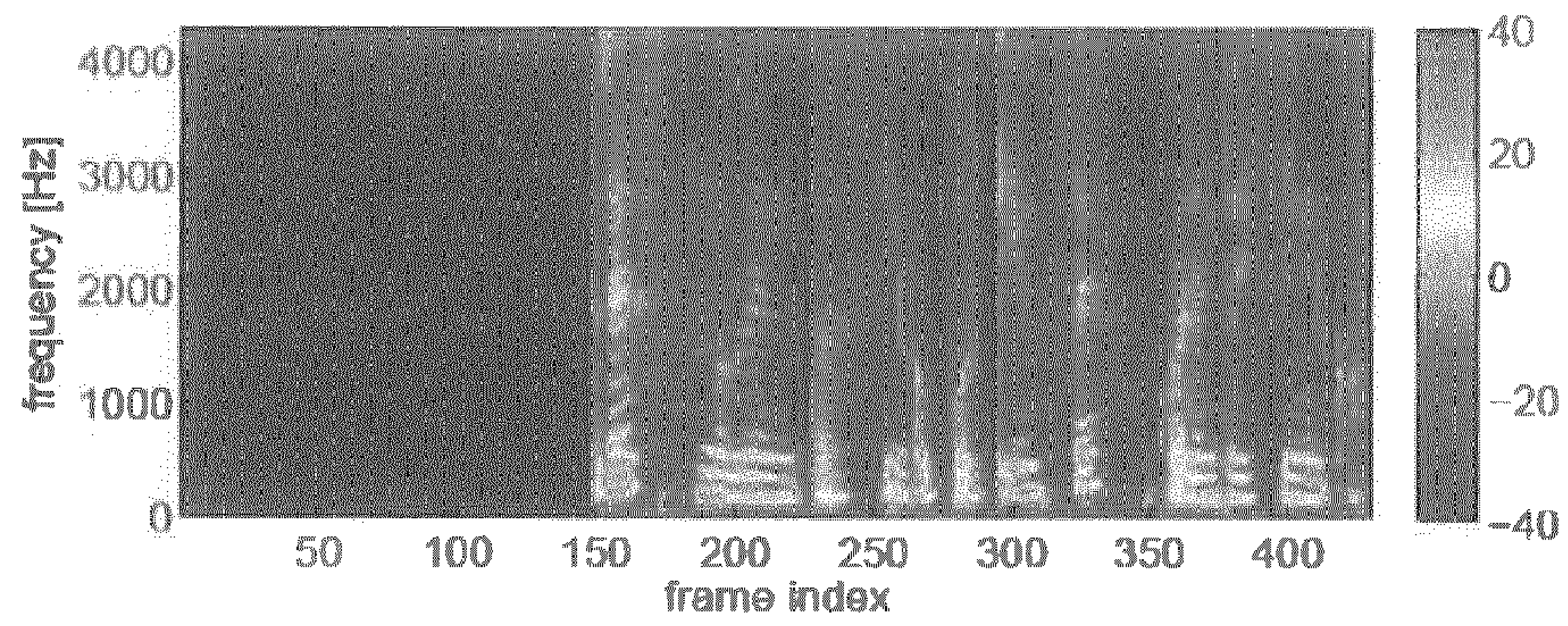


FIG 11A

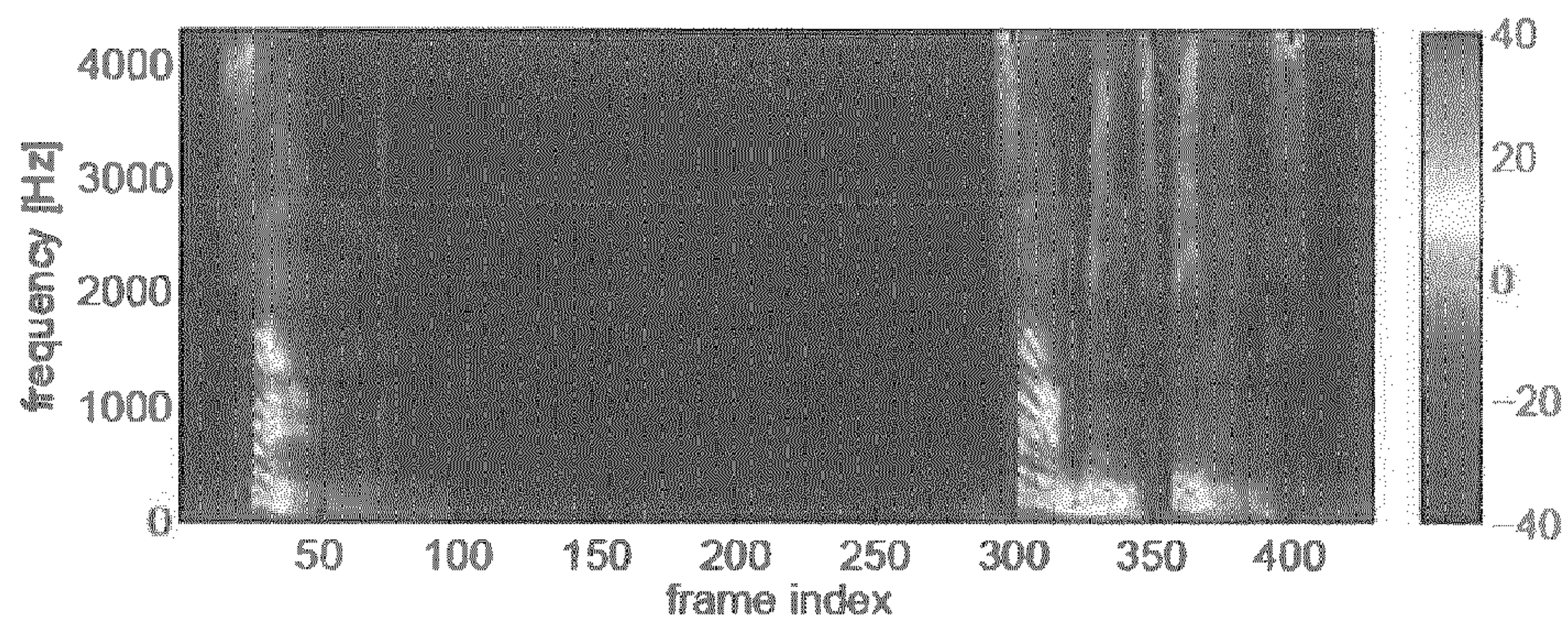


FIG 11B

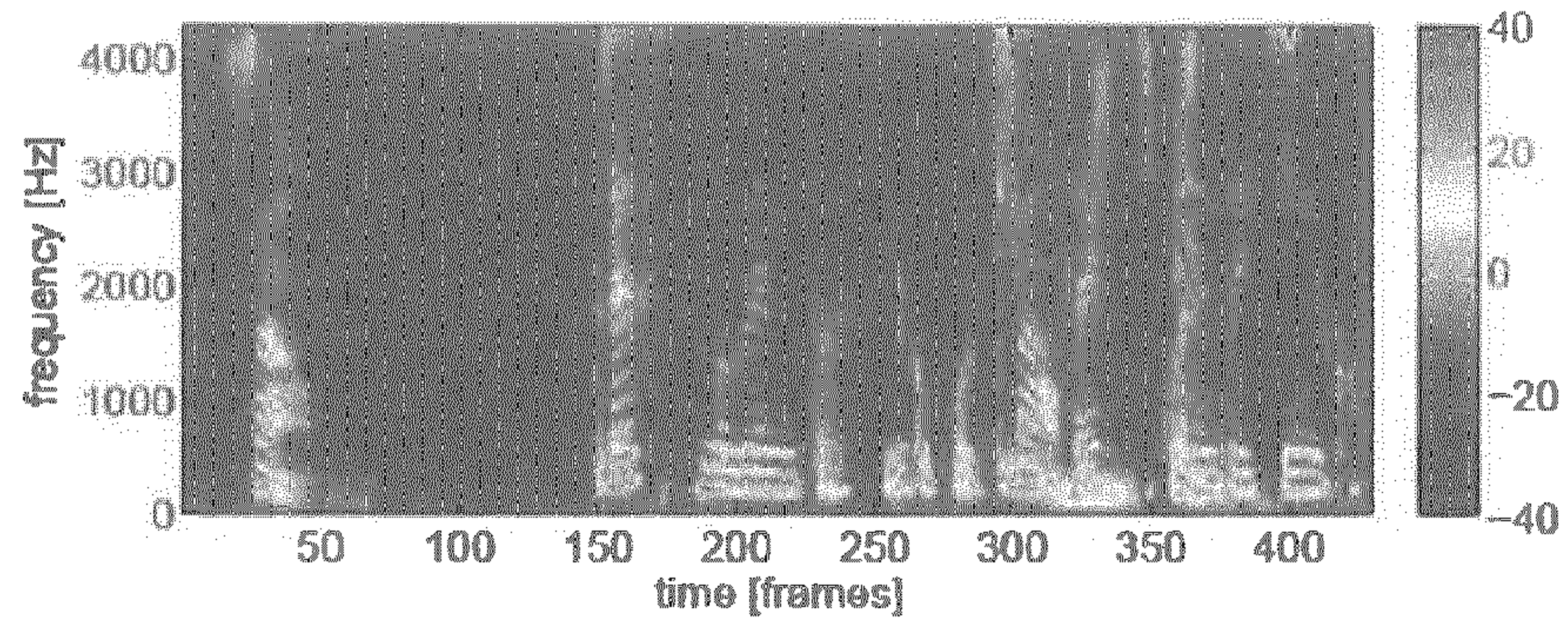


FIG 11C

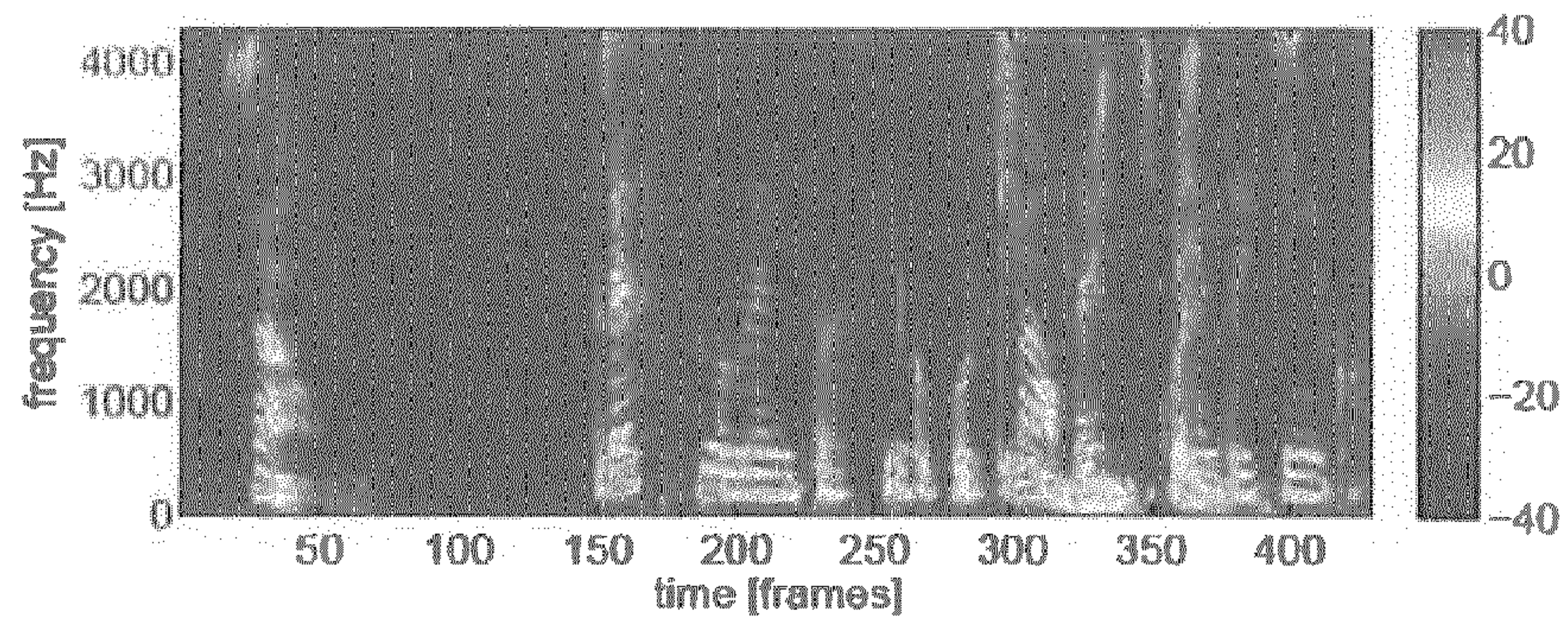


FIG 11D

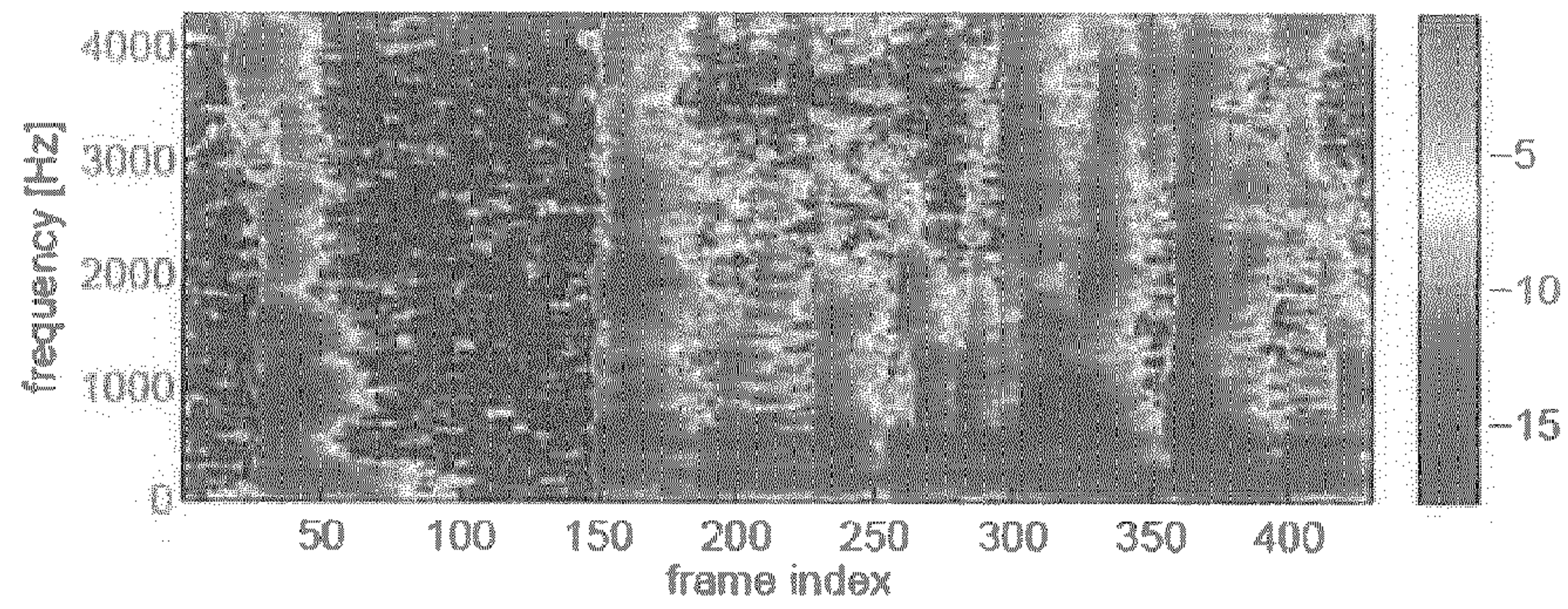


FIG 12A

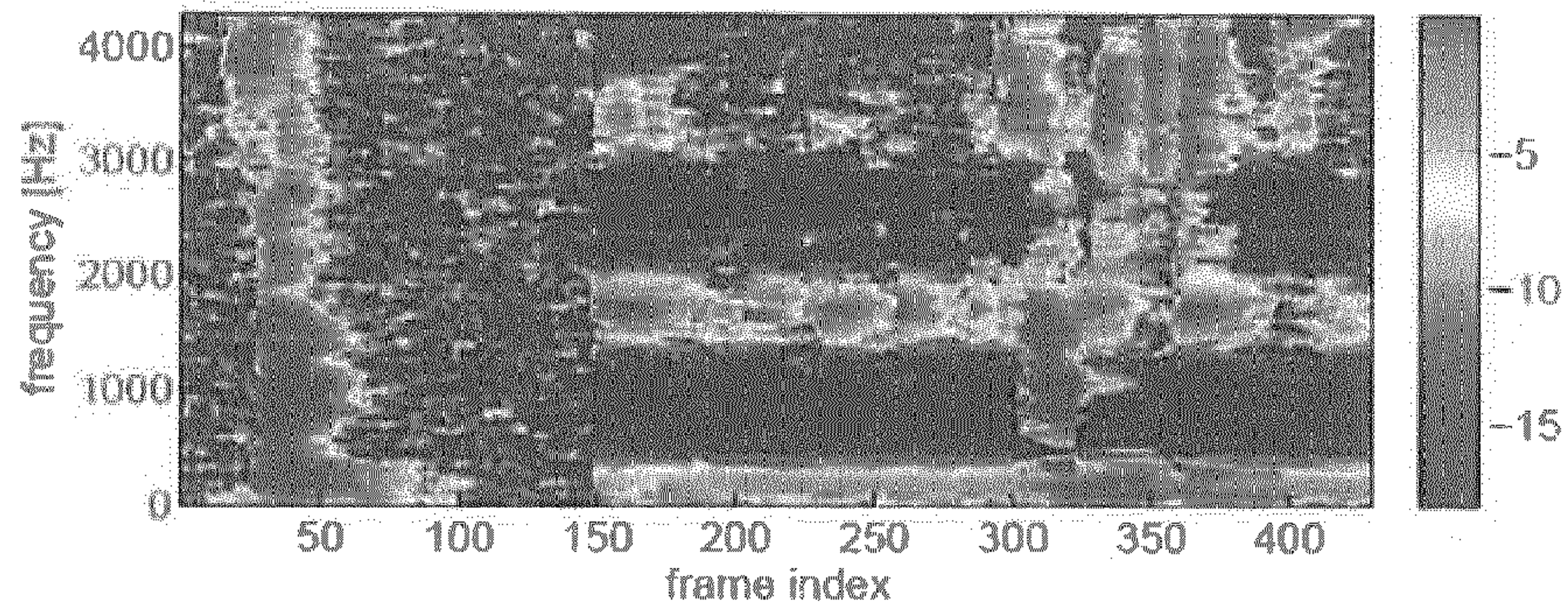


FIG 12B

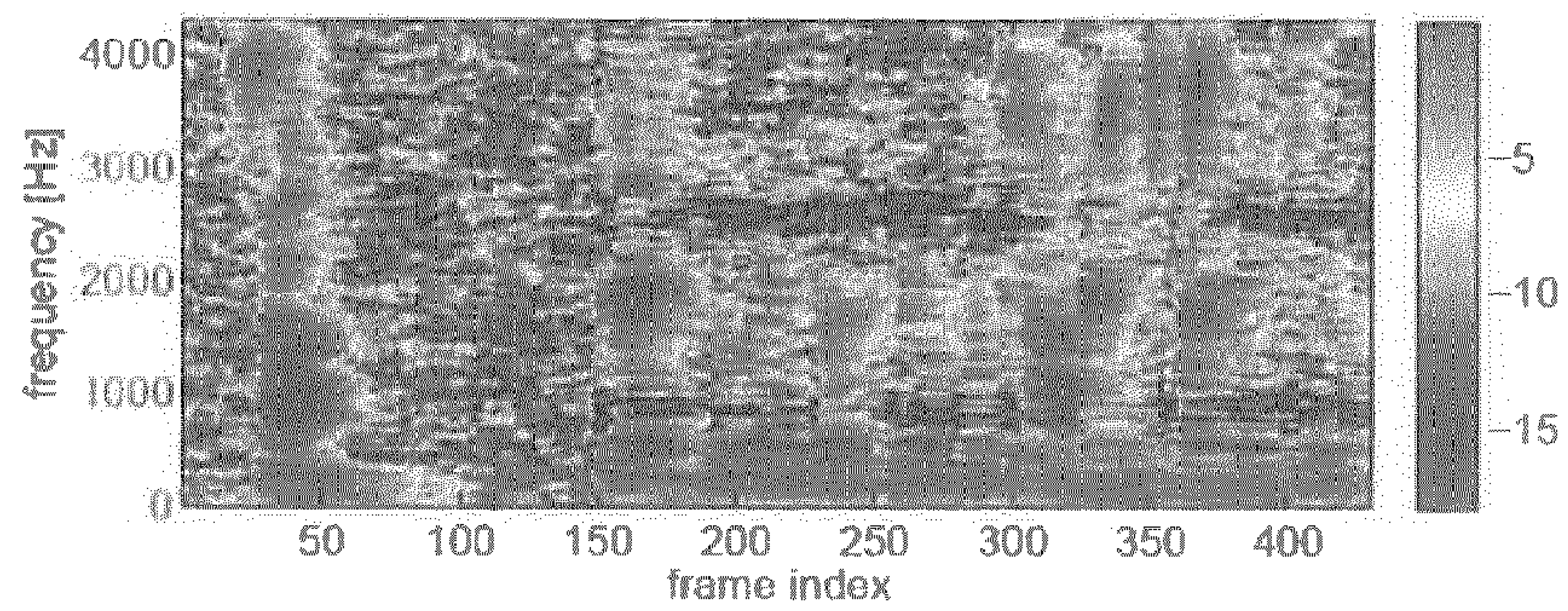


FIG 12C

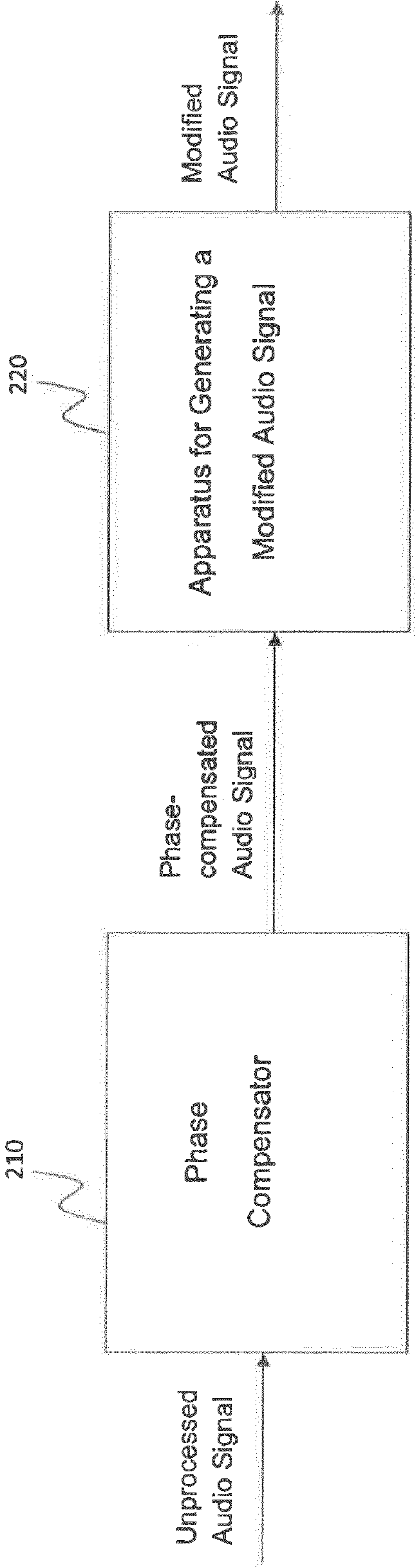


FIG 13

