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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 April 2018 (19.04.2018)

(10) International Publication Number
WO 2018/071387 A1

(51) International Patent Classification:
G10L 21/0216 (2013.01) *G10L 25/30* (2013.01)

(21) International Application Number:
PCT/US2017/055887

(22) International Filing Date:
10 October 2017 (10.10.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
15/290,685 11 October 2016 (11.10.2016) US
15/583,012 01 May 2017 (01.05.2017) US

(71) Applicant: CIRRUS LOGIC INTERNATIONAL SEMICONDUCTOR, LTD. [GB/GB]; 7B Nightingale Way, Quartermile, Edinburgh Lothian EH3 9EG (GB).

(72) Inventor; and

(71) Applicant *(for US only)*: **EBENEZER, Samuel Pon Varma** [US/US]; 4630 S. Lakeshore Drive, #189, Tempe, Arizona 85282 (US).

(74) Agent: PREWITT, Brian K. et al.; JACKSON WALKER L.L.P., 100 Congress Avenue, Suite 1100, Austin, Texas 78701 (US).

(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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, 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,

(54) Title: DETECTION OF ACOUSTIC IMPULSE EVENTS IN VOICE APPLICATIONS USING A NEURAL NETWORK

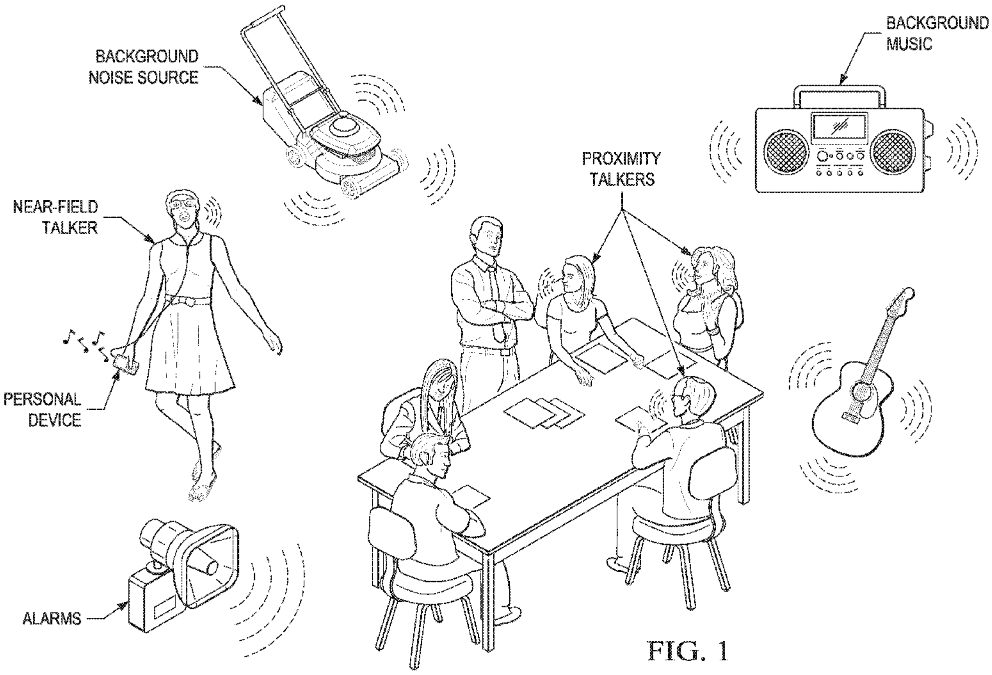


FIG. 1

(57) Abstract: In accordance with embodiments of the present disclosure, an integrated circuit for implementing at least a portion of an audio device may include an audio output configured to reproduce audio information by generating an audio output signal for communication to at least one transducer of the audio device, a microphone input configured to receive an input signal indicative of ambient sound external to the audio device, and a processor configured to implement an impulsive noise detector. The impulsive noise detector may comprise a plurality of processing blocks for determining a feature vector based on characteristics of the input signal and a neural network for determining based on the feature vector whether the impulsive event comprises a speech event or a noise event.

WO 2018/071387 A1

[Continued on next page]

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.

(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, ST, 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).

Published:

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

DETECTION OF ACOUSTIC IMPULSE EVENTS IN VOICE APPLICATIONS USING A NEURAL NETWORK

Related Applications

This application claims priority to U.S. Pat. Appl. No. 15/583,012, filed May 1, 2017, which application claims continuation-in-part priority to U.S. Pat. Appl. No. 15/290,685, filed October 11, 2016, and entitled "Detection of Acoustic Impulse Events in Voice Applications," which is incorporated herein by reference.

Technical Field

The field of representative embodiments of this disclosure relates to methods, apparatuses, and implementations concerning or relating to voice applications in an audio device. Applications include detection of acoustic impulsive noise events using a neural network based on the harmonic and sparse spectral nature of speech.

Background

Voice activity detection (VAD), also known as speech activity detection or speech detection, is a technique used in speech processing in which the presence or absence of human speech is detected. VAD may be used in a variety of applications, including noise suppressors, background noise estimators, adaptive beamformers, dynamic beam steering, always-on voice detection, and conversation-based playback management. In many of such applications, high-energy and transient background noises that are often present in an environment are impulsive in nature. Many traditional VADs rely on changes in signal level on a full-band or sub-band basis and thus often detect such impulsive noise as speech, as a signal envelope of an impulsive noise is often similar to that of speech. In addition, in many cases, an impulsive noise spectrum averaged over various impulsive noise occurrences and an averaged speech spectrum may not be significantly different. Accordingly, in such systems, impulsive noise may be detected as speech, which may deteriorate system performance. For example, in a beamsteering application, false detection of an impulse noise as speech may result in steering a “look” direction of the beam-steering system in an incorrect direction even though an individual speaking is not moving relative to the audio device.

Summary

In accordance with the teachings of the present disclosure, one or more disadvantages and problems associated with existing approaches to voice activity detection may be reduced or eliminated.

In accordance with embodiments of the present disclosure, an integrated circuit for implementing at least a portion of an audio device may include an audio output configured to reproduce audio information by generating an audio output signal for communication to at least one transducer of the audio device, a microphone input configured to receive an input signal indicative of ambient sound external to the audio device, and a processor configured to implement an impulsive noise detector. The impulsive noise detector may comprise a plurality of processing blocks for determining a feature vector based on characteristics of the input signal and a neural network for determining, based on the feature vector, whether the impulsive event comprises a speech event or a noise event.

In accordance with these and other embodiments of the present disclosure, a method for impulsive noise detection may include receiving an input signal indicative of ambient sound external to an audio device, determining a feature vector based on characteristics of the input signal, using a neural network to determine based on the feature vector whether the impulsive event comprises a speech event or a noise event, and reproducing audio information by generating an audio output signal for communication to at least one transducer of an audio device based on the input signal and the determination of whether the impulsive event comprises a speech event or a noise event.

Technical advantages of the present disclosure may be readily apparent to one of ordinary skill in the art from the figures, description, and claims included herein. The objects and

advantages of the embodiments will be realized and achieved at least by the elements, features, and combinations particularly pointed out in the claims.

It is to be understood that both the foregoing general description and the following detailed description are examples and explanatory and are not restrictive of the claims set forth in this disclosure.

Brief Description of the Drawings

A more complete understanding of the example, present embodiments and certain advantages thereof may be acquired by referring to the following description taken in conjunction with the accompanying drawings, in which like reference numbers indicate like features, and wherein:

Figure 1 illustrates an example of a use case scenario wherein various detectors may be used in conjunction with a playback management system to enhance a user experience, in accordance with embodiments of the present disclosure;

Figure 2 illustrates an example playback management system, in accordance with embodiments of the present disclosure;

Figure 3 illustrates an example steered response power based beamsteering system, in accordance with embodiments of the present disclosure;

Figure 4 illustrates an example adaptive beamformer, in accordance with embodiments of the present disclosure;

Figure 5 illustrates a block diagram of an example impulsive noise detector, in accordance with embodiments of the present disclosure;

Figure 6 illustrates a block diagram of an example feature vector pre-processing block, in accordance with embodiments of the present disclosure;

Figure 7 illustrates a block diagram of an example neural network for detecting acoustic impulse events, in accordance with embodiments of the present disclosure;

Figure 8 illustrates a block diagram of an example neural network for detecting acoustic impulse events with example weights associated with each of the synaptic connections between neurons in different layers and example corresponding non-linear functions used at each neuron, in accordance with embodiments of the present disclosure;

Figure 9 illustrates a block diagram of an example approach to augmenting neural network training data, in accordance with embodiments of the present disclosure; and

Figure 10 illustrates a block diagram of an example impulsive noise detector, in accordance with embodiments of the present disclosure.

Detailed Description

In accordance with embodiments of this disclosure, an automatic playback management framework may use one or more audio event detectors. Such audio event detectors for an audio device may include a near-field detector that may detect when sounds in the near-field of the audio device are detected, such as when a user of the audio device (e.g., a user that is wearing or otherwise using the audio device) speaks, a proximity detector that may detect when sounds in proximity to the audio device are detected, such as when another person in proximity to the user of the audio device speaks, and a tonal alarm detector that detects acoustic alarms that may have been originated in the vicinity of the audio device. Figure 1 illustrates an example of a use case scenario wherein such detectors may be used in conjunction with a playback management system to enhance a user experience, in accordance with embodiments of the present disclosure.

Figure 2 illustrates an example playback management system that modifies a playback signal based on a decision from an event detector 2, in accordance with embodiments of the present disclosure. Signal processing functionality in a processor 7 may comprise an acoustic echo canceller 1 that may cancel an acoustic echo that is received at microphones 9 due to an echo coupling between an output audio transducer 8 (e.g., loudspeaker) and microphones 9. The echo reduced signal may be communicated to event detector 2 which may detect one or more various ambient events, including without limitation a near-field event (e.g., including but not limited to speech from a user of an audio device) detected by near-field detector 3, a proximity event (e.g., including but not limited to speech or other ambient sound other than near-field sound) detected by proximity detector 4, and/or a tonal alarm event detected by alarm detector 5.

If an audio event is detected, an event-based playback control 6 may modify a characteristic of audio information (shown as “playback content” in Figure 2) reproduced to output audio transducer 8. Audio information may include any information that may be reproduced at output audio transducer 8, including without limitation, downlink speech associated with a telephonic conversation received via a communication network (e.g., a cellular network) and/or internal audio from an internal audio source (e.g., music file, video file, etc.).

As shown in Figure 2, near-field detector 3 may include a voice activity detector 11 which may be utilized by near-field detector 3 to detect near-field events. Voice activity detector 11 may include any suitable system, device, or apparatus configured to perform speech processing to detect the presence or absence of human speech. In accordance with such processing, voice activity detector 11 may include an impulsive noise detector 12. In operation, as described in greater detail below, impulsive noise detector 12 may predict an occurrence of a signal burst event of an input signal indicative of ambient sound external to an audio device (e.g., a signal induced by sound pressure on one or more microphones 9) to determine whether the signal burst event comprises a speech event or a noise event.

As shown in Figure 2, proximity detector 4 may include a voice activity detector 13 which may be utilized by proximity detector 4 to detect events in proximity with an audio device. Similar to voice activity detector 11, voice activity detector 13 may include any suitable system, device, or apparatus configured to perform speech processing to detect the presence or absence of human speech. In accordance with such processing, voice activity detector 13 may include an impulsive noise detector 14. Similar to impulsive noise detector 12, impulsive noise detector 14 may predict an occurrence of a signal burst event of an input signal indicative of ambient sound external to an audio device (e.g., a signal induced by sound pressure on one or more microphones 9) to determine whether the signal burst event comprises a speech event or a noise event. In some embodiments, processor 7 may include a single voice activity detector having a single impulsive noise detector leveraged by both of near-field detector 3 and proximity detector 4 in performing their functionality.

Figure 3 illustrates an example steered response power-based beamsteering system 30, in accordance with embodiments of the present disclosure. Steered response power-based beamsteering system 30 may operate by implementing multiple beamformers 33 (e.g., delay-and-sum and/or filter-and-sum beamformers) each with a different look direction such that the entire bank of beamformers 33 will cover the desired field of interest. The beamwidth of each beamformer 33 may depend on a microphone array aperture length. An output power from each beamformer 33 may be computed, and a beamformer 33 having a maximum output power may be switched to an output path 34 by a steered-response power-based beam selector 35. Switching of beam selector 35 may be constrained by a voice activity detector 31 having an impulsive noise detector 32 such that the output power is measured by beam selector 35 only when speech is detected, thus preventing beam selector 35 from rapidly switching between multiple beamformers 33 by responding to spatially non-stationary background impulsive noises.

Figure 4 illustrates an example adaptive beamformer 40, in accordance with embodiments of the present disclosure. Adaptive beamformer 40 may comprise any system, device, or apparatus capable of adapting to changing noise conditions based on the received data. In general, an adaptive beamformer may achieve higher noise cancellation or interference suppression compared to fixed beamformers. As shown in Figure 4, adaptive beamformer 40 is implemented as a generalized side lobe canceller (GSC). Accordingly, adaptive beamformer 40 may comprise a fixed beamformer 43, blocking matrix 44, and a multiple-input adaptive noise canceller 45 comprising an adaptive filter 46. If adaptive filter 46 were to adapt at all times, it may train to speech leakage also causing speech distortion during a subtraction stage 47. To increase robustness of adaptive beamformer 40, a voice activity detector 41 having an impulsive noise detector 42 may communicate a control signal to adaptive filter 46 to disable training or adaptation in the presence of speech. In such implementations, voice activity detector 41 may control a noise estimation period wherein background noise is not estimated whenever speech is present. Similarly, the robustness of a GSC to speech leakage may be further improved by using an adaptive blocking matrix, the control for which may include an improved voice activity detector with an impulsive noise detector, as described in U.S. Pat. No. 9,607,603 entitled "Adaptive Block Matrix Using Pre-Whitening for Adaptive Beam Forming."

Figure 5 illustrates a block diagram of an example impulsive noise detector 50, in accordance with embodiments of the present disclosure. In some embodiments, impulsive noise detector 50 may implement one or more of impulsive noise detector 12, impulsive noise detector 14, impulsive noise detector 32, and impulsive noise detector 42. Impulsive noise detector 50 may comprise any suitable system, device, or apparatus configured to exploit the harmonic nature of speech to distinguish impulsive noise from speech, as described in greater detail below.

As shown in Figure 5, impulsive noise detector 50 may comprise a plurality of processing blocks (e.g., sudden onset calculation block 53, harmonicity calculation block 55, a harmonic product spectrum block 56, a harmonic flatness measure block 57, a spectral flatness measure (SFM) block 58, and a SFM swing block 59) for determining a feature vector based on characteristics of an input audio signal x[n], a feature vector pre-processing block 60 for pre-processing the feature vector to generate a dimension-reduced feature vector, and a neural network 70 for determining based on the feature vector (e.g., using the dimension-reduced feature vector) whether an impulsive noise event has occurred, as described in greater detail below.

Sudden onset calculation block 53 may comprise any system, device, or apparatus configured to exploit sudden changes in a signal level of input audio signal x[n] in order to generate statistics indicative of an existence or an absence of a forthcoming signal burst. For example, samples of input audio signal x[n] may first be grouped into overlapping frame samples and the energy of each frame computed. Sudden onset calculation block 53 may calculate the energy (shown as "energy" in Figure 5) of a frame as:

$$E[l] = \sum_{n=1}^{N} x^2[n, l]$$

where N is the total number of samples in a frame, l is the frame index, and a predetermined percentage (e.g., 25%) of overlapping is used to generate each frame. Such energy measure statistic generated by sudden onset calculation block 53 may be communicated as a portion of the feature vector communicated to feature vector pre-processing block 60 and neural network 70.

Further, sudden onset calculation block 53 may calculate a sudden onset statistic (shown as "suddenOnSetStat" in Figure 5) which exploits sudden changes in the signal level of input audio signal x[n] to predict a forthcoming signal burst. A normalized frame energy may be calculated as:

$$\hat{E}[m,l] = \frac{E[m]}{\max_{\forall m} E[m] - \min_{\forall m} E[m]}$$

where m = l, l-1, l-2, …, l-L+1 and L is a size of the frame energy history buffer. The denominator in this normalization step may represent a dynamic range of frame energy over the current and past (L-1) frames. Sudden onset calculation block 53 may compute a sudden onset statistic as:

$$\gamma_{os}[l] = \frac{\max_{\forall m'} \hat{E}[m',l]}{\hat{E}[l,l]}$$

where m' = l-1, l-2, …, l-L+1. One of skill in the art may note that the maximum is computed only over the past (L-1) frames. Therefore, if a sudden acoustic event appears in the environment, the frame energy at the onset of the event may be high and the maximum energy over the past (L-1) frames may be smaller than the maximum value. Therefore, the ratio of these two values may be small during the onset. Accordingly, the frame size should be such that the past (L-1) frames do not contain energy corresponding to the signal burst.

Because sudden onset calculation block 53 detects signal level fluctuations, the sudden onset statistic may also indicate the presence of sudden speech bursts. For example, the sudden onset statistic may indicate the presence of sudden speech bursts every time a speech event appears after a period of silence. Accordingly, impulsive noise detector 50 cannot rely solely on sudden onset calculation block 53 to accurately detect an impulsive noise, and other statistics described below (e.g., harmonicity, harmonic product spectrum flatness measure, spectral flatness measure, and/or spectral flatness measure swing of audio input signal x[n]) may exploit the harmonic and sparse nature of an instantaneous speech spectrum of audio input signal x[n] to determine if a signal onset is caused by speech or impulsive noise.

In order to extract spectral information of audio input signal x[n] in order to determine values of such parameters, impulsive noise detector 50 may convert audio input signal x[n] from the time domain to the frequency domain by means of a discrete Fourier transform (DFT) 54. DFT 54 may buffer, overlap, window, and convert audio input signal x[n] to the frequency domain as:

$$X[k,l] = \sum_{n=0}^{N-1} w[n]x[n,l]e^{\frac{-j2\pi nk}{N}}, \quad k = 0, 1, \dots, N-1,$$

where $w[n]$ is a windowing function, $x[n,l]$ is a buffered and overlapped input signal frame, N is a size of the DFT size and k is a frequency bin index. The overlap may be fixed at any suitable percentage (e.g., 25%).

To determine harmonicity, harmonicity calculation block 55 may compute total power in a frame as:

$$E_x[l] = \sum_{k \in \mathcal{K}} |X[k,l]|^2$$

where $\mathcal{K}$ is a set of all frequency bin indices corresponding to the spectral range of interest. Harmonicity calculation block 55 may calculate a harmonic power as:

$$E_H[p,l] = \sum_{m=1}^{N_h} |X[mp,l]|^2, \quad p \in \mathcal{P}$$

where N_h is a number of harmonics, m is a harmonic order, and $\mathcal{P}$ is a set of all frequency bin indices corresponding to an expected pitch frequency range. The expected pitch frequency range may be set to any suitable range (e.g., 100-500Hz). A harmonicity at a given frequency may be defined as a ratio of the harmonic power to the total energy without the harmonic power and harmonicity calculation block 55 may calculate harmonicity as:

$$H[p,l] = \frac{E_H[p,l]}{E_x[l] - E_H[p,l]}.$$

For clean speech signals, harmonicity may have a maximum at the pitch frequency. Because an impulsive noise spectrum may be less sparse than a speech spectrum, harmonicity for impulsive noises may be small. Thus, a harmonicity calculation block 55 may output a harmonicity-based test statistic (shown as "harmonicity" in Figure 5) formulated as:

$$\gamma_{Harm}[l] = \max_{\forall p \in \mathcal{P}} H[p,l].$$

In many instances, most of impulsive noises corresponding to transient acoustic events tend to have more energy at lower frequencies. Moreover, the spectrum may also typically be less sparse at these lower frequencies. On the other hand, a spectrum corresponding to voiced speech also has more low-frequency energy. However, in most instances, a speech spectrum has more sparsity than impulsive noises. Therefore, one can examine the flatness of the spectrum at these lower frequencies as a deterministic factor. Accordingly, SFM block 58 may calculate a sub-band spectral flatness measure (shown as "sfm" in Figure 5) computed as:

$$\gamma_{SFM}[l] = \frac{\prod_{k=N_L}^{N_H}[|X[k,l]|^2]^{1/N_B}}{\frac{1}{N_B}\sum_{k=N_L}^{N_H}|X[k,l]|^2}$$

where $N_B = N_H - N_L + 1$, N_H and N_L are the spectral bin indices corresponding to low- and high-frequency band edges respectively, of a sub-band. The sub-band frequency range may be of any suitable range (e.g., 500-1500 Hz).

An ability to differentiate speech from impulsive noise based on harmonicity may degrade when non-impulsive background noise is also present in an acoustic environment. Under such conditions, harmonic product spectrum block 56 may provide more robust harmonicity information. Harmonic product spectrum block 56 may calculate a harmonic product spectrum as:

$$G[p,l] = \prod_{m=1}^{N_h} |X[mp,l]|^2, \quad p \in \mathcal{P}$$

where N_h and $\mathcal{P}$ are defined above with respect to the calculation of harmonicity. The harmonic product spectrum tends to have a high value at the pitch frequency since the pitch frequency harmonics are accumulated constructively, while at other frequencies, the harmonics are accumulated destructively. Therefore, the harmonic product spectrum is a sparse spectrum for speech, and it is less sparse for impulsive noise because the noise energy in impulsive noise distributes evenly across all frequencies. Therefore, a flatness of the harmonic product spectrum may be used as a differentiating factor. Harmonic flatness measure block 57 may compute a flatness measure of the harmonic product spectrum (shown as “hpsSfm” in Figure 5) as:

$$\gamma_{HPS-SFM}[l] = \frac{\prod_{p \in \mathcal{P}}[|G[p,l]|^2]^{1/N_{\mathcal{P}}}}{\frac{1}{N_{\mathcal{P}}}\sum_{p \in \mathcal{P}}|G[p,l]|^2}$$

where $N_{\mathcal{P}}$ is the number of spectral bins in the pitch frequency range.

An impulsive noise spectrum may exhibit spectral stationarity over a short period of time (e.g., 300 - 500 ms), whereas a speech spectrum may vary over time due to spectral modulation of pitch harmonics. Once a signal burst onset is detected, SFM swing block 59 may capture such non-stationarity information by tracking spectral flatness measures from multiple sub-bands over a period of time and estimate the variation of the weighted and cumulative flatness measure over the same period. For example, SFM swing block 59 may track a cumulative SFM over a period of time and may calculate a difference between the maximum and the minimum cumulative SFM value over the same duration, such difference representing a flatness measure swing. The flatness measure swing value may generally be small for impulsive noises because the spectral content of such signals may be wideband in nature and may tend to be stationary for a short interval of time. The value of the flatness measure swing value may be higher for speech signals because spectral content of speech signals may vary faster than impulsive noises. SFM swing block 59 may calculate the flatness measure swing by first computing the cumulative spectral flatness measure as:

$$\rho_{SFM}[l] = \sum_{i=1}^{N_S} \alpha(i) \left\{ \frac{\prod_{k=N_L(i)}^{N_H(i)} [|X[k,l]|^2]^{1/N_B(i)}}{\frac{1}{N_B(i)} \sum_{k=N_L(i)}^{N_H(i)} |X[k,l]|^2} \right\}$$

where $N_B(i) = N_H(i) - N_L(i) + 1$, i is a sub-band number, N_s is a number of sub-bands, $\alpha(i)$ is a sub-band weighting factor, $N_H(i)$ and $N_L(i)$ are spectral bin indices corresponding to the low- and high-frequency band edges, respectively of i^{th} sub-band. Any suitable sub-band ranges may be employed (e.g., 500-1500 Hz, 1500-2750 Hz, and 2750-3500 Hz). SFM swing block 59 may then smooth the cumulative spectral flatness measure as:

$$\mu_{SFM}[l] = \beta * \mu_{SFM}[l-1] + (1-\beta)\rho_{SFM}[l]$$

where β is the exponential averaging smoothing coefficient. SFM swing block 59 may obtain the spectral flatness measure swing by computing a difference between a maximum and a minimum spectral flatness measure value over the most-recent M frames. Thus, SFM swing block 59 may generate a spectral flatness measure swing-based test statistic (shown as "sfmSwing" in Figure 5) defined as:

$$\gamma_{SFM-Swing}[l] = \max_{\forall m=l,l-1,l-M+1} \mu_{SFM}[m] - \min_{\forall m=l,l-1,l-M+1} \mu_{SFM}[m].$$

Together, the statistics described above may define a feature vector which may be used as an input by neural network 70 (e.g., after pre-processing by feature vector pre-processing block 60). For example, the feature vector corresponding to an l^{th} frame may be given by:

$$v\text{: } \left[E[l] \quad \gamma_{os}[l] \quad \gamma_{Harm}[l] \quad \gamma_{SFM}[l] \quad \gamma_{HPS-SFM}[l] \quad \gamma_{SFMSwing}[l]\right]^T.$$

Because durations of impulsive noises may be longer than a frame size under consideration, impulsive noise detection performance of neural network 70 may be improved if feature vectors from multiple frames are used in the decision process. Figure 6 illustrates a block diagram of an example feature vector pre-processing block 60, in accordance with embodiments of the present disclosure. As described below, feature vector pre-processing block 60 may augment a current feature set with the feature set of previous frames, and then further process such augmented feature set to generate a dimension-reduced feature vector to facilitate efficient computation by neural network 70.

As shown in Figure 6, frames of a feature vector may be stored in a long-term memory 62 such that a feature augmentation block 64 may augment a current frame of the feature vector with a suitable number of past frames of the feature vector to generate an augmented feature vector that may be given as:

$$\boldsymbol{v}^r\text{: } \left[E[l] \cdots \gamma_{SFMSwing}[l]\; E[l-1] \cdots \gamma_{SFMSwing}[l-1]\; E[l-L] \cdots \gamma_{SFMSwing}[l-L]\right]^T.$$

Because a dynamic range of each statistic in the feature vector may be different, standardization block 66 may normalize the statistics with respect to each other in order to reduce or eliminate feature bias. For example, in some embodiments, standardization block 66 may scale each statistic to have a zero mean and unit variance in a statistical sense. Thus, a raw statistic may be scaled as:

$$\vartheta_i^s = \frac{\vartheta_i^r - \mu_i^r}{\sigma_i^r}$$

where $\vartheta_i^r[l]$ and $\vartheta_i^s[l]$ are the raw and scaled i^{th} feature respectively, μ_i^r and σ_i^r are the mean and standard deviation of i^{th} feature and they are obtained using the sample mean/variance estimate as:

$$\mu_i^r = \frac{1}{K}\sum_{k=1}^{K} \vartheta_i^{r,k}$$

$$\sigma_i^r = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\vartheta_i^{r,k} - \mu_i^r\right)^2}$$

where K is the number of training samples, $\vartheta_i^{r,k}$ corresponds to the i^{th} raw feature from k^{th} training sample.

Because augmentation of the feature vector with past frames increases the feature dimension of the feature vector, the number of weights needed in neural network 70 to process the feature vector may be affected. Accordingly, feature dimension reduction block 68 may reduce the feature dimension of the normalized augmented feature vector to generate a dimension-reduced feature vector for more efficient processing by neural network 70. For example, the temporal correlation that exists between successive frames may increase feature redundancy, and feature dimension reduction such as principle component analysis may be employed by feature dimension reduction block 68 to compress the feature set. To illustrate, the principal components may be estimated by first estimating a covariance matrix of the feature vector as:

$$\boldsymbol{C}_{\vartheta} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\vartheta}^{s,k^{T}} \boldsymbol{\vartheta}^{s,k}$$

where $\boldsymbol{\vartheta}^{s,k}$ is the k^{th} standardized training sample. The covariance matrix may be decomposed using an eigen analysis as:

$$\boldsymbol{C}_{\vartheta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{T}$$

where $\boldsymbol{D}$ is a diagonal matrix containing the eigen values sorted in descending order, $\boldsymbol{U}$ is the matrix containing the unitary eigen vectors and the superscript T represents the matrix transpose. The principal components may be obtained by selecting the eigen vectors corresponding to not insignificant eigen values. These eigen values may be chosen by either comparing the fraction of an eigen value to the maximum value or by selecting all the eigen values whose cumulative value encompasses a certain percentage of the signal energy. For example, a threshold of 99% of the signal energy may be used as a threshold for the eigen vector selection. Thus, the feature dimension may be reduced by applying a linear transform on the feature vector as:

$$\boldsymbol{\vartheta} = \begin{bmatrix} \boldsymbol{u}_1^T \\ \vdots \\ \boldsymbol{u}_P^T \end{bmatrix} \boldsymbol{\vartheta}^{s}$$

where $\boldsymbol{u}_1$ corresponds to the eigen vector corresponding to the largest eigen value, P is the number of principal components, $\boldsymbol{\vartheta}^{s}$ is the standardized feature vector and $\boldsymbol{\vartheta}$ is the dimension reduced feature vector.

Figure 7 illustrates a block diagram of an example neural network 70 for detecting acoustic impulse events, in accordance with embodiments of the present disclosure. As shown in Figure 7, neural network 70 may comprise an input layer 72 of the statistics described above (e.g., as pre-processed by feature vector pre-processing block 60 to generate a dimension-reduced feature vector) and a plurality of hidden layers 74 comprising neurons 76 coupled together via synapses which may perform neural processing to determine if an impulsive noise event is present, to generate a single-output output layer 78 that may output a signal *indDet* indicative of a determination of the presence of impulsive noise.

Figure 8 illustrates a block diagram of an example neural network 70 for detecting acoustic impulse events with example weights associated with each of the synaptic connections between neurons 76 in different layers and example corresponding non-linear functions used at each neuron 76, in accordance with embodiments of the present disclosure. The various weights of the neural network may be computed by training the neural network with a set of training samples. A desired output d[n] corresponding to output signal *indDet* for the training samples may be generated by labeling each frame of data as either impulsive noise or not. The neural weights set forth in Figure 8 may be calculated by reducing an error e[n] between output signal *indDet* and the desired output d[n]. Error e[n] may be back propagated to input layer 72 and the various weights may be updated recursively using a back propagation algorithm as follows:

For output layer 78:

$$w_{m1}^{(L)}[n+1] = w_{m1}^{(L)}[n] + \mu\delta_m^{(L)}[n]y_m^{(L)}[n],\ \ m = 1,\cdots,\ M(L)$$

$$\delta_m^{(L)}[n] = e[n]\varphi'\left\{v_m^{(L)}[n]\right\}$$

$$e[n] = y_1^{(L+1)}[n] - d[n]$$

$$\varphi'\left\{v_m^{(L)}[n]\right\} = \frac{ae^{-av_m^{(L)}[n]}}{\left[1 + e^{-av_m^{(L)}[n]}\right]^2}$$

For hidden layers 74:

$$w_{mi}^{(l)}[n+1] = w_{mi}^{(l)}[n] + \mu\delta_m^{(l)}[n]y_m^{(l)}[n],$$

$$m = 1,\cdots,\ M(l),\ \ i = 1,\cdots,\ M(l+1),\ \ l = 1,\cdots,\ L-1$$

$$\delta_m^{(l)}[n] = \varphi'\left\{v_m^{(l)}[n]\right\}\sum_{k=1}^{M(l)}\delta_k^{(l+1)}[n]w_{mk}^{(l)}[n]$$

where $M(l)$ is the number of nodes in layer l. In some embodiments, neural network 70 may be trained using a scaled conjugate gradient back propagation update method that has smaller memory requirements at the expense of more iteration required for convergence of weights. A cross-entropy error may be used in the cost function. The cross-entropy error may be given by:

$$C_{\varepsilon}=\frac{1}{K}\sum_{k=1}^{K} d_k \log(y_k^{L+1}) + (1-d_k)\log(1-y_k^{L+1})$$

where y_k^{L+1} is the neural network output corresponding to k^{th} training sample.

In order to reduce generalization error, neural network 70 must be trained with a large volume of training data. The training data may be increased by augmenting the available training samples through amplitude, time, and frequency scaling, as shown in Figure 9. Such data augmentation may help neural network 70 to be invariant to amplitude, time and frequency scaling. This may be advantageous for an impulsive noise detector because impulsive noises are random in nature and the amplitude, and duration and spectrum of impulsive noises can vary depending on the stimulus source. For example, if $\{(\boldsymbol{v}_1,d_1), (\boldsymbol{v}_2,d_2), \ldots (\boldsymbol{v}_K,d_K)\}$ represents K training sample sets, then an augmented training set obtained by amplitude scaling the input signal, $x^g[n] = gx[n]$ may be represented as $\{ (\boldsymbol{v}_1,d_1), (\boldsymbol{v}_2,d_2), \ldots (\boldsymbol{v}_K,d_K),(\boldsymbol{v}_1^g, d_1), (\boldsymbol{v}_2^g, d_2), \ldots (\boldsymbol{v}_K^g,d_K)\}$.

Figure 10 illustrates a block diagram of another example impulsive noise detector 50A, in accordance with embodiments of the present disclosure. Impulsive noise detector 50A may be similar to impulsive noise detector 50 of Figure 5, with a feature vector used by neural network 70 enhanced with mel cepstral coefficients of the input audio signal x[n], which may be extracted by mel cepstral coefficients block 52. Inclusion of mel cepstral coefficients in the feature vector may increase the performance of neural network 70, at the cost of increased processing resources required due to the increased number of neurons needed due to the increase in feature dimension that results.

It should be understood — especially by those having ordinary skill in the art with the benefit of this disclosure — that the various operations described herein, particularly in connection with the figures, may be implemented by other circuitry or other hardware components. The order in which each operation of a given method is performed may be changed, and various elements of the systems illustrated herein may be added, reordered, combined, omitted, modified, etc. It is intended that this disclosure embrace all such modifications and changes and, accordingly, the above description should be regarded in an illustrative rather than a restrictive sense.

Similarly, although this disclosure makes reference to specific embodiments, certain modifications and changes can be made to those embodiments without departing from the scope and coverage of this disclosure. Moreover, any benefits, advantages, or solutions to problems that are described herein with regard to specific embodiments are not intended to be construed as a critical, required, or essential feature or element.

Further embodiments likewise, with the benefit of this disclosure, will be apparent to those having ordinary skill in the art, and such embodiments should be deemed as being encompassed herein.

WHAT IS CLAIMED IS:

1. An integrated circuit for implementing at least a portion of an audio device, comprising:

an audio output configured to reproduce audio information by generating an audio output signal for communication to at least one transducer of the audio device;

a microphone input configured to receive an input signal indicative of ambient sound external to the audio device; and

a processor configured to implement an impulsive noise detector comprising:

a plurality of processing blocks for determining a feature vector based on characteristics of the input signal; and

a neural network for determining based on the feature vector whether an impulsive event comprises a speech event or a noise event.

2. The integrated circuit of Claim 1, wherein the processor is further configured to modify a characteristic associated with the audio information in response to detection of a noise event.

3. The integrated circuit of Claim 2, wherein the characteristic comprises one or more of an amplitude of the audio information and spectral content of the audio information.

4. The integrated circuit of Claim 2, wherein the characteristic comprises at least one coefficient of a voice-based processing algorithm including at least one of a noise suppressor, a background noise estimator, an adaptive beamformer, dynamic beam steering, always-on voice, and a conversation-based playback management system.

5. The integrated circuit of Claim 1, wherein the feature vector comprises statistics indicative of harmonicity, sparsity, and degree of temporal modulation of a signal spectrum of the input signal to determine whether the impulsive event comprises a speech event or a noise event.

6. The integrated circuit of Claim 1, wherein the feature vector comprises a statistic indicative of an acoustic energy present in the input signal.

7. The integrated circuit of Claim 1, wherein the feature vector comprises a statistic indicative of an occurrence of a signal burst event of the input signal.

8. The integrated circuit of Claim 1, wherein the feature vector comprises a statistic indicative of mel cepstral coefficients of the input signal.

9. The integrated circuit of Claim 1, further comprising a pre-processing block interfaced between the plurality of processing blocks and the neural network and configured to:

augment the feature vector with at least one previous frame of the feature vector to generate an augmented feature vector;

normalize statistics of the feature vector with respect to each other; and

reduce feature dimensions of the statistics of the feature vector prior to training by the neural network.

10. A method for impulsive noise detection comprising:

receiving an input signal indicative of ambient sound external to an audio device;

determining a feature vector based on characteristics of the input signal;

using a neural network to determine based on the feature vector whether an impulsive event comprises a speech event or a noise event; and

reproducing audio information by generating an audio output signal for communication to at least one transducer of an audio device based on the input signal and the determination of whether the impulsive event comprises a speech event or a noise event.

11. The method of Claim 10, further comprising modifying a characteristic associated with the audio information in response to detection of a noise event.

12. The method of Claim 11, wherein the characteristic comprises one or more of an amplitude of the audio information and spectral content of the audio information.

13. The method of Claim 11, wherein the characteristic comprises at least one coefficient of a voice-based processing algorithm including at least one of a noise suppressor, a background noise estimator, an adaptive beamformer, dynamic beam steering, always-on voice, and a conversation-based playback management system.

14. The method of Claim 10, wherein the feature vector comprises statistics indicative of harmonicity, sparsity, and degree of temporal modulation of a signal spectrum of the input signal to determine whether the impulsive event comprises a speech event or a noise event.

15. The method of Claim 10, wherein the feature vector comprises a statistic indicative of an acoustic energy present in the input signal.

16. The method of Claim 10, wherein the feature vector comprises a statistic indicative of an occurrence of a signal burst event of the input signal.

17. The method of Claim 10, wherein the feature vector comprises a statistic indicative of mel cepstral coefficients of the input signal.

18. The method of Claim 10, further comprising pre-processing the feature vector prior to using the neural network, wherein pre-processing comprises:

augmenting the feature vector with at least one previous frame of the feature vector to generate an augmented feature vector;

normalizing statistics of the feature vector with respect to each other; and

reducing feature dimensions of the statistics of the feature vector prior to training by the neural network.

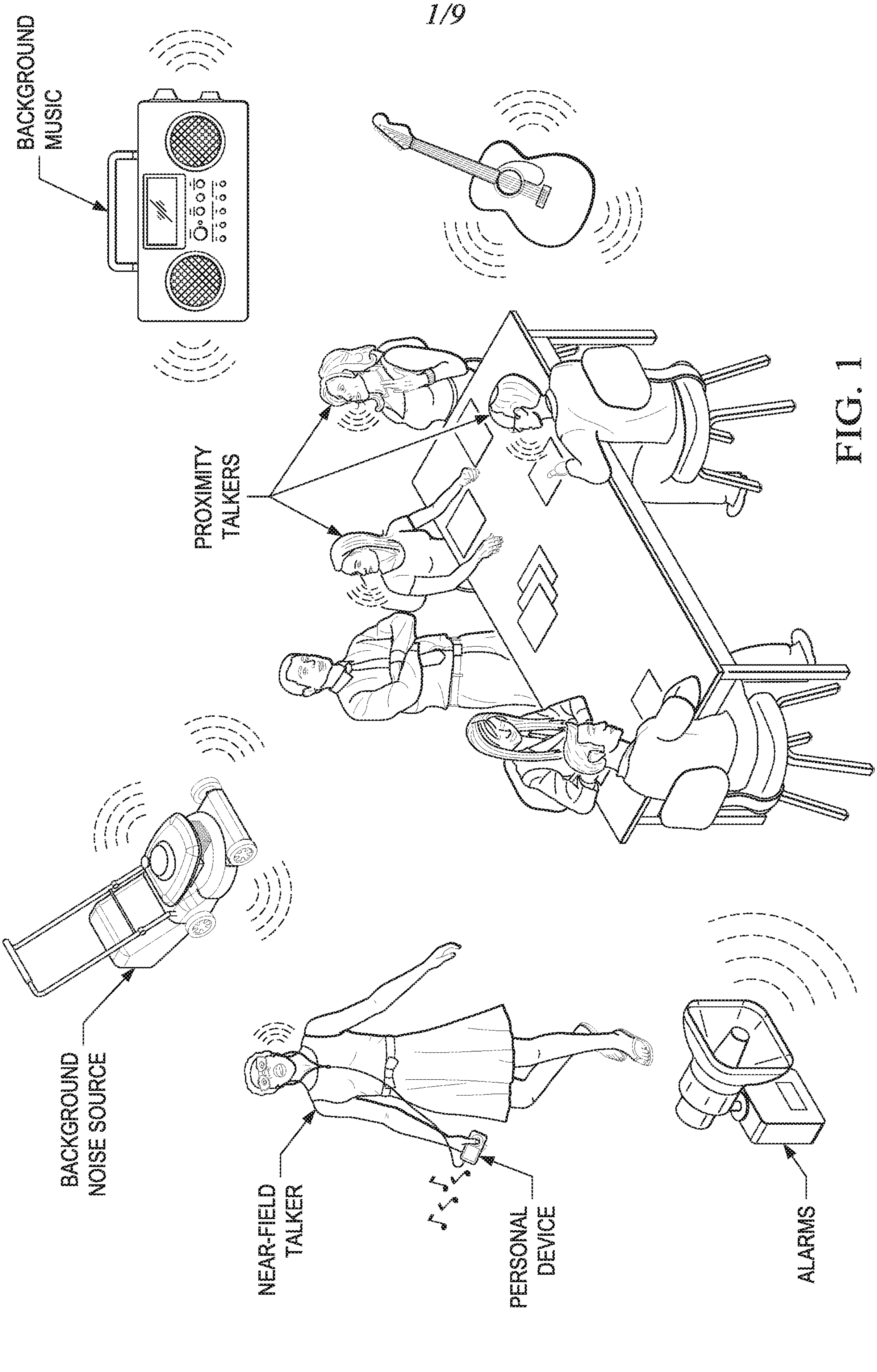

BACKGROUND
MUSIC
PROXIMITY
TALKERS
BACKGROUND
NOISE SOURCE
NEAR-FIELD
TALKER
PERSONAL
DEVICE
ALARMS

FIG. 1

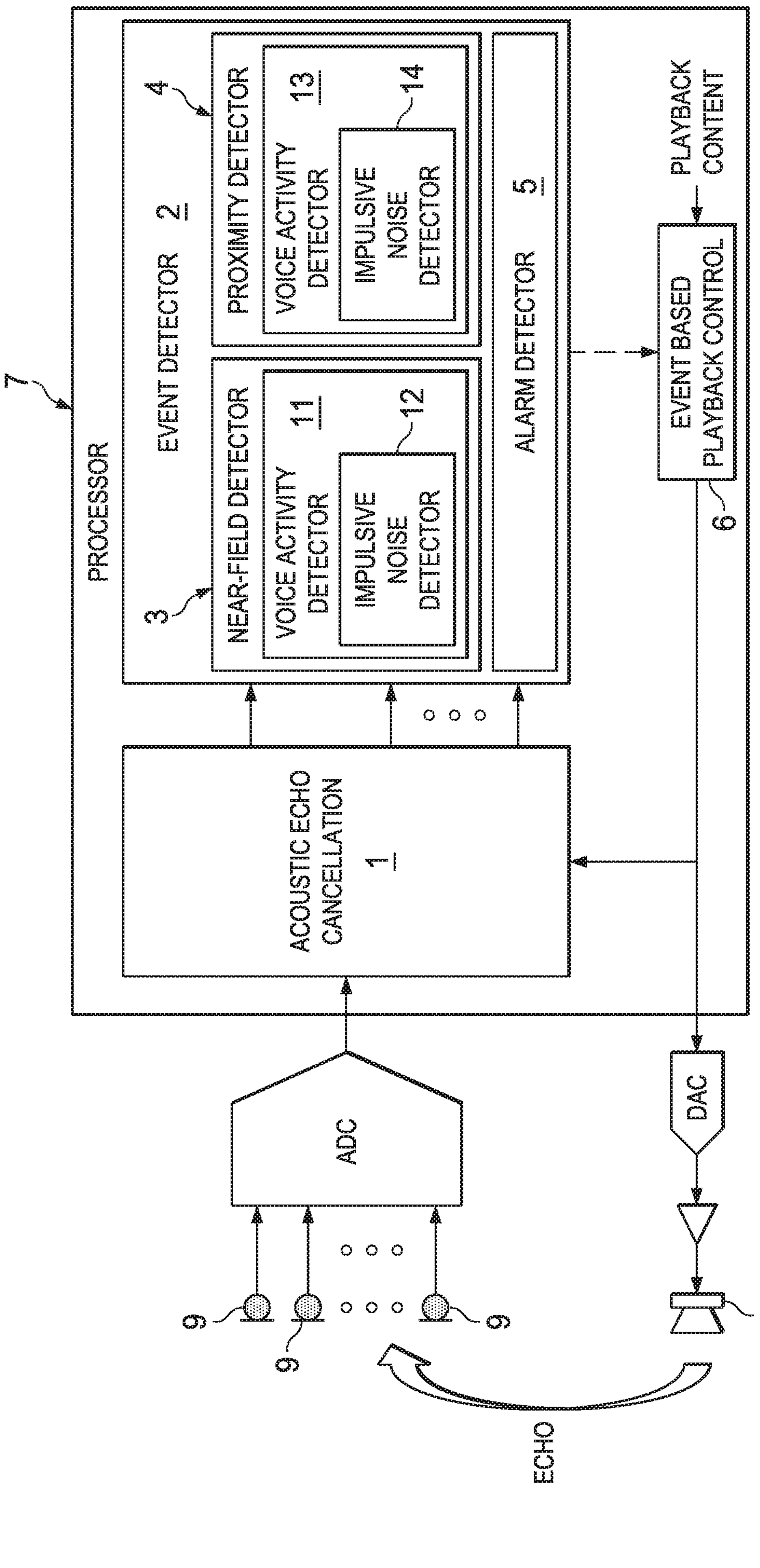

PROCESSOR
7
EVENT DETECTOR 2
NEAR-FIELD DETECTOR
3
VOICE ACTIVITY DETECTOR 11
IMPULSIVE NOISE DETECTOR
12
PROXIMITY DETECTOR
4
VOICE ACTIVITY DETECTOR 13
IMPULSIVE NOISE DETECTOR
14
ALARM DETECTOR 5
EVENT BASED PLAYBACK CONTROL
6
PLAYBACK CONTENT
ACOUSTIC ECHO CANCELLATION 1
ADC
DAC
9
9
9
8
ECHO

FIG. 2

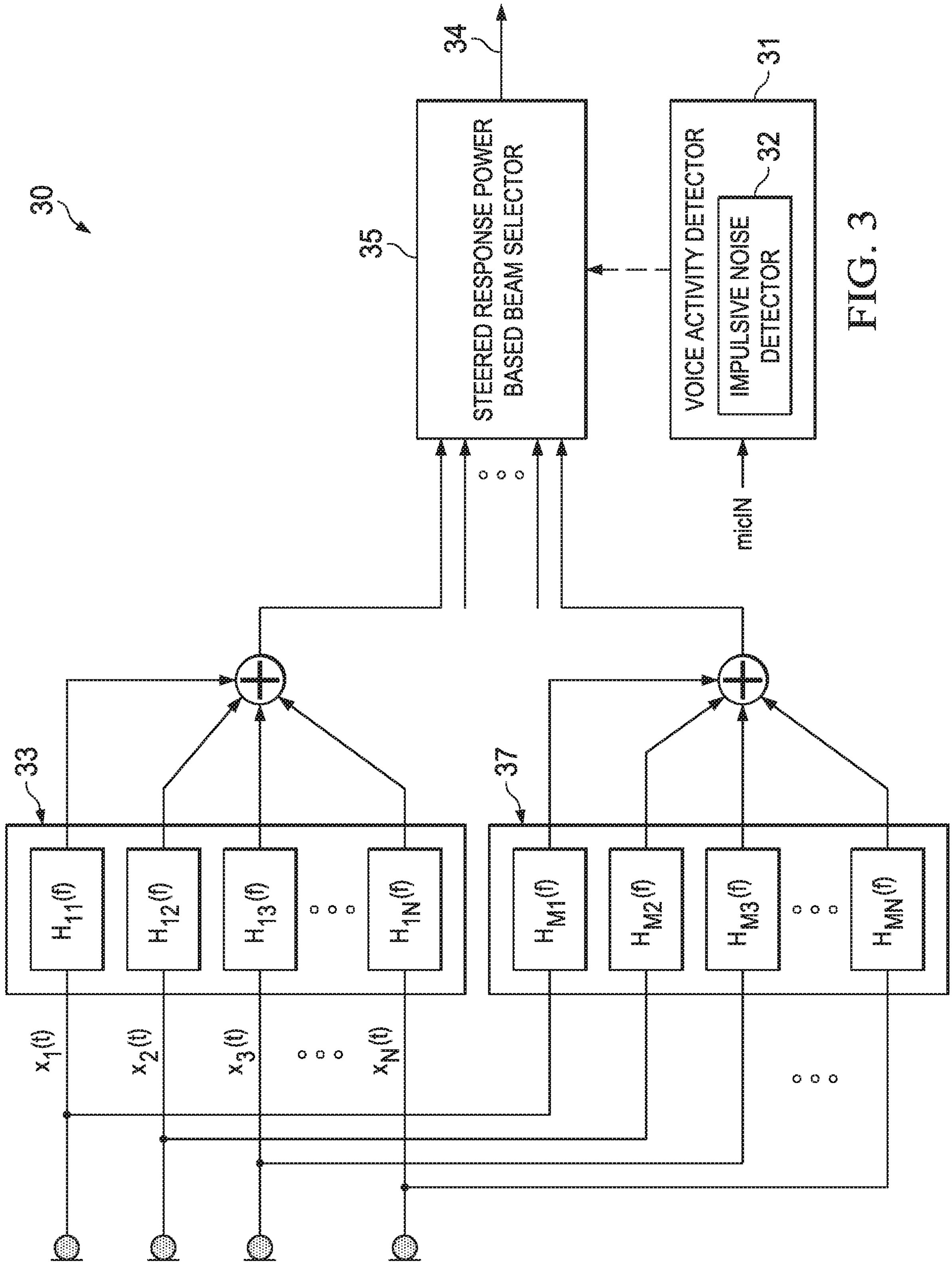
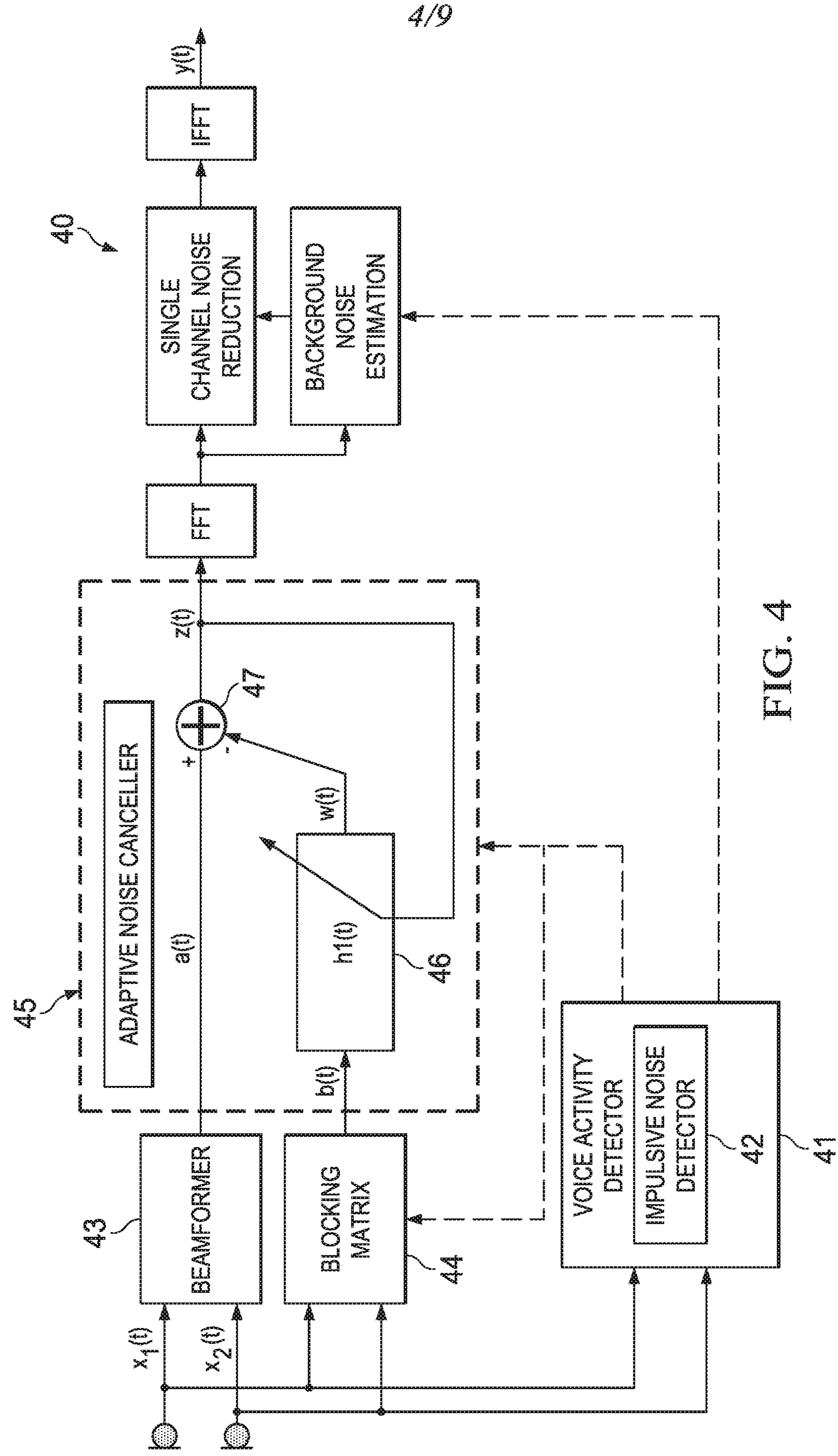

30
34
35
STEERED RESPONSE POWER BASED BEAM SELECTOR
31
VOICE ACTIVITY DETECTOR
32
IMPULSIVE NOISE DETECTOR
micN
33
37
$H_{11}(f)$
$H_{12}(f)$
$H_{13}(f)$
$H_{1N}(f)$
$H_{M1}(f)$
$H_{M2}(f)$
$H_{M3}(f)$
$H_{MN}(f)$
$x_1(t)$
$x_2(t)$
$x_3(t)$
$x_N(t)$

FIG. 3



y(t)
IFFT
40
SINGLE CHANNEL NOISE REDUCTION
BACKGROUND NOISE ESTIMATION
FFT
z(t)
47
+
-
45
ADAPTIVE NOISE CANCELLER
a(t)
w(t)
h1(t)
46
b(t)
43
BEAMFORMER
BLOCKING MATRIX
44
$x_1(t)$
$x_2(t)$
VOICE ACTIVITY DETECTOR
IMPULSIVE NOISE DETECTOR
42
41

FIG. 4

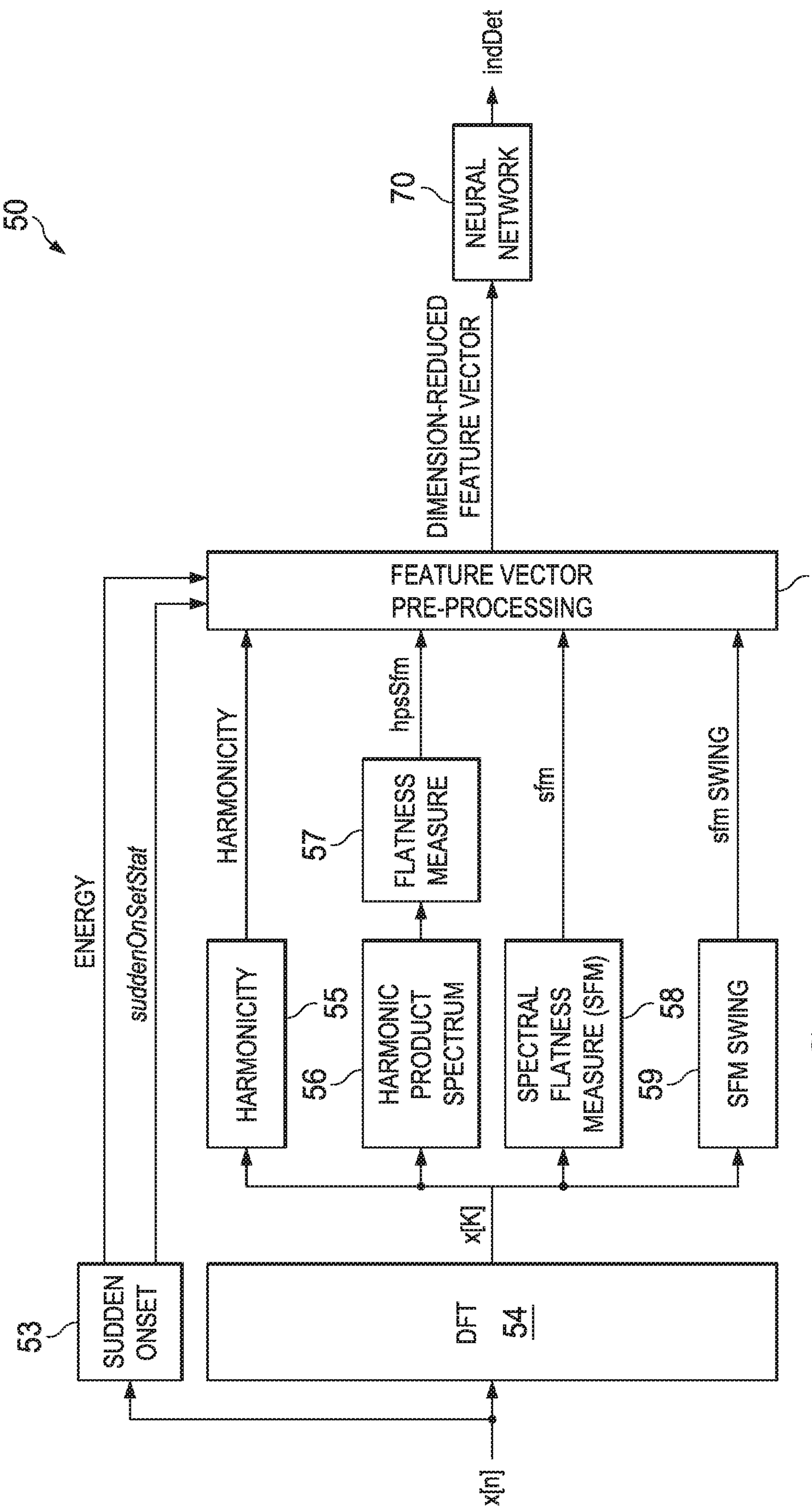

50
x[n]
SUDDEN ONSET
53
DFT 54
x[K]
ENERGY
suddenOnSetStat
HARMONICITY
55
HARMONIC PRODUCT SPECTRUM
56
SPECTRAL FLATNESS MEASURE (SFM)
58
SFM SWING
59
FLATNESS MEASURE
57
HARMONICITY
hpsSfm
sfm
sfm SWING
FEATURE VECTOR PRE-PROCESSING
60
DIMENSION-REDUCED FEATURE VECTOR
NEURAL NETWORK
70
indDet

FIG. 5

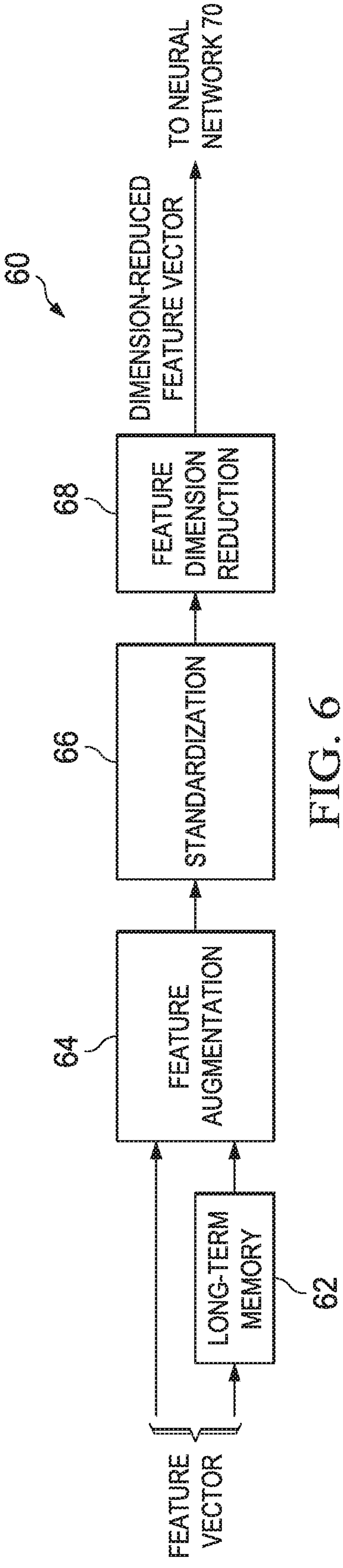

60
FEATURE VECTOR
LONG-TERM MEMORY
62
64
FEATURE AUGMENTATION
66
STANDARDIZATION
68
FEATURE DIMENSION REDUCTION
DIMENSION-REDUCED FEATURE VECTOR
TO NEURAL NETWORK 70

FIG. 6

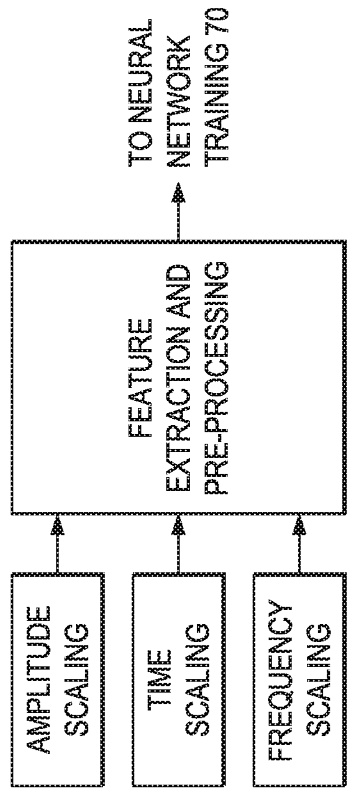

AMPLITUDE SCALING
TIME SCALING
FREQUENCY SCALING
FEATURE EXTRACTION AND PRE-PROCESSING
TO NEURAL NETWORK TRAINING 70

FIG. 9

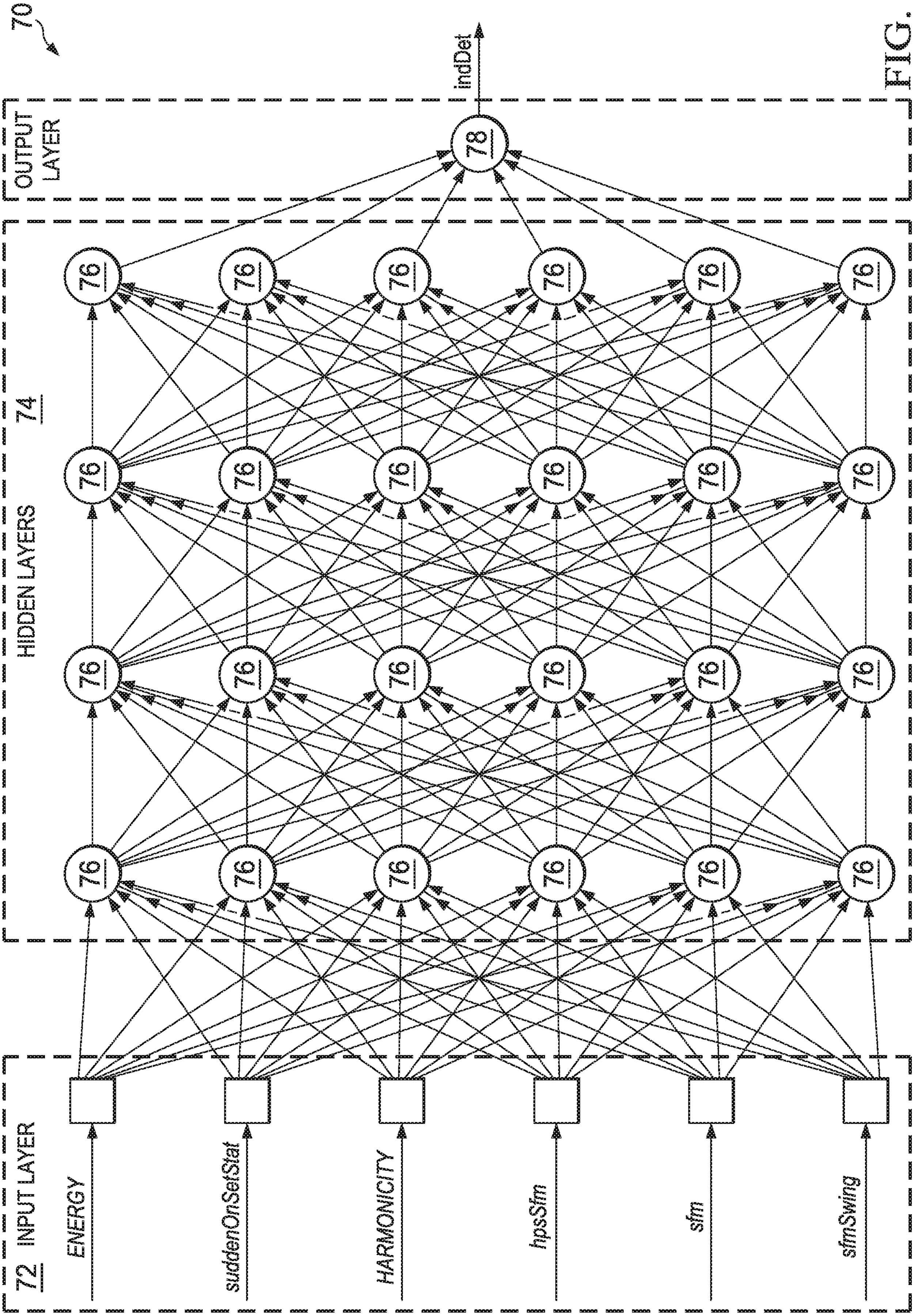

70
72 INPUT LAYER
ENERGY
suddenOnSetStat
HARMONICITY
hpsSfm
sfm
sfmSwing
HIDDEN LAYERS 74
76
OUTPUT LAYER
78
indDet

FIG. 7

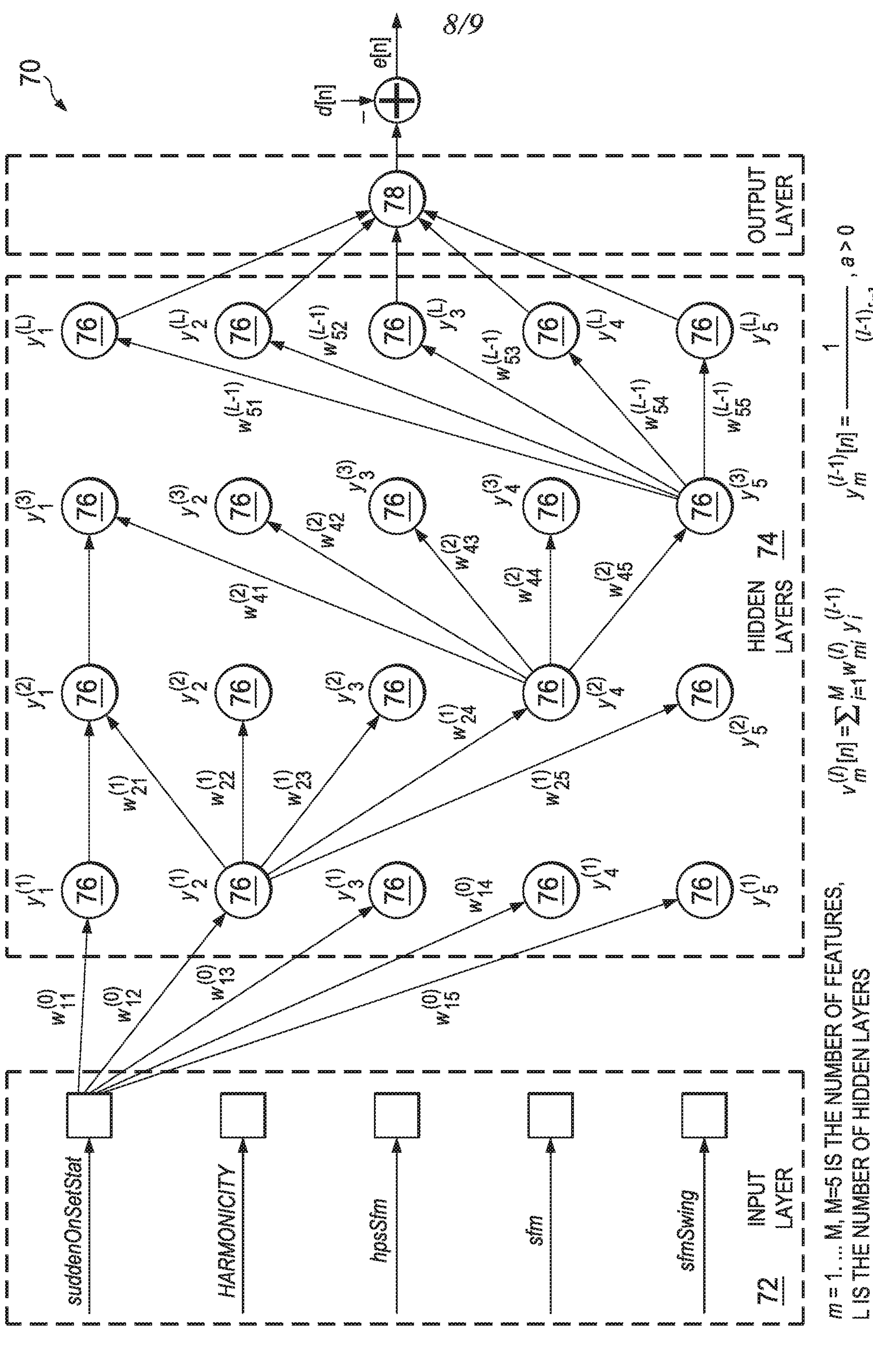

70
INPUT LAYER
72
suddenOnSetStat
HARMONICITY
hpsSfm
sfm
sfmSwing
HIDDEN LAYERS
74
76
OUTPUT LAYER
78
d[n]
e[n]
$w^{(0)}_{11}$ $w^{(0)}_{12}$ $w^{(0)}_{13}$ $w^{(0)}_{14}$ $w^{(0)}_{15}$
$w^{(1)}_{21}$ $w^{(1)}_{22}$ $w^{(1)}_{23}$ $w^{(1)}_{24}$ $w^{(1)}_{25}$
$w^{(2)}_{41}$ $w^{(2)}_{42}$ $w^{(2)}_{43}$ $w^{(2)}_{44}$ $w^{(2)}_{45}$
$w^{(L-1)}_{51}$ $w^{(L-1)}_{52}$ $w^{(L-1)}_{53}$ $w^{(L-1)}_{54}$ $w^{(L-1)}_{55}$
$y^{(1)}_1$ $y^{(1)}_2$ $y^{(1)}_3$ $y^{(1)}_4$ $y^{(1)}_5$
$y^{(2)}_1$ $y^{(2)}_2$ $y^{(2)}_3$ $y^{(2)}_4$ $y^{(2)}_5$
$y^{(3)}_1$ $y^{(3)}_2$ $y^{(3)}_3$ $y^{(3)}_4$ $y^{(3)}_5$
$y^{(L)}_1$ $y^{(L)}_2$ $y^{(L)}_3$ $y^{(L)}_4$ $y^{(L)}_5$
$v^{(l)}_m[n] = \sum_{i=1}^{M} w^{(l)}_{mi} y^{(l-1)}_i$
$y^{(l-1)}_m[n] = \frac{1}{1 + e^{-av^{(l-1)}_m[n]}}, a > 0$
m = 1, ... M, M=5 IS THE NUMBER OF FEATURES,
L IS THE NUMBER OF HIDDEN LAYERS

FIG. 8

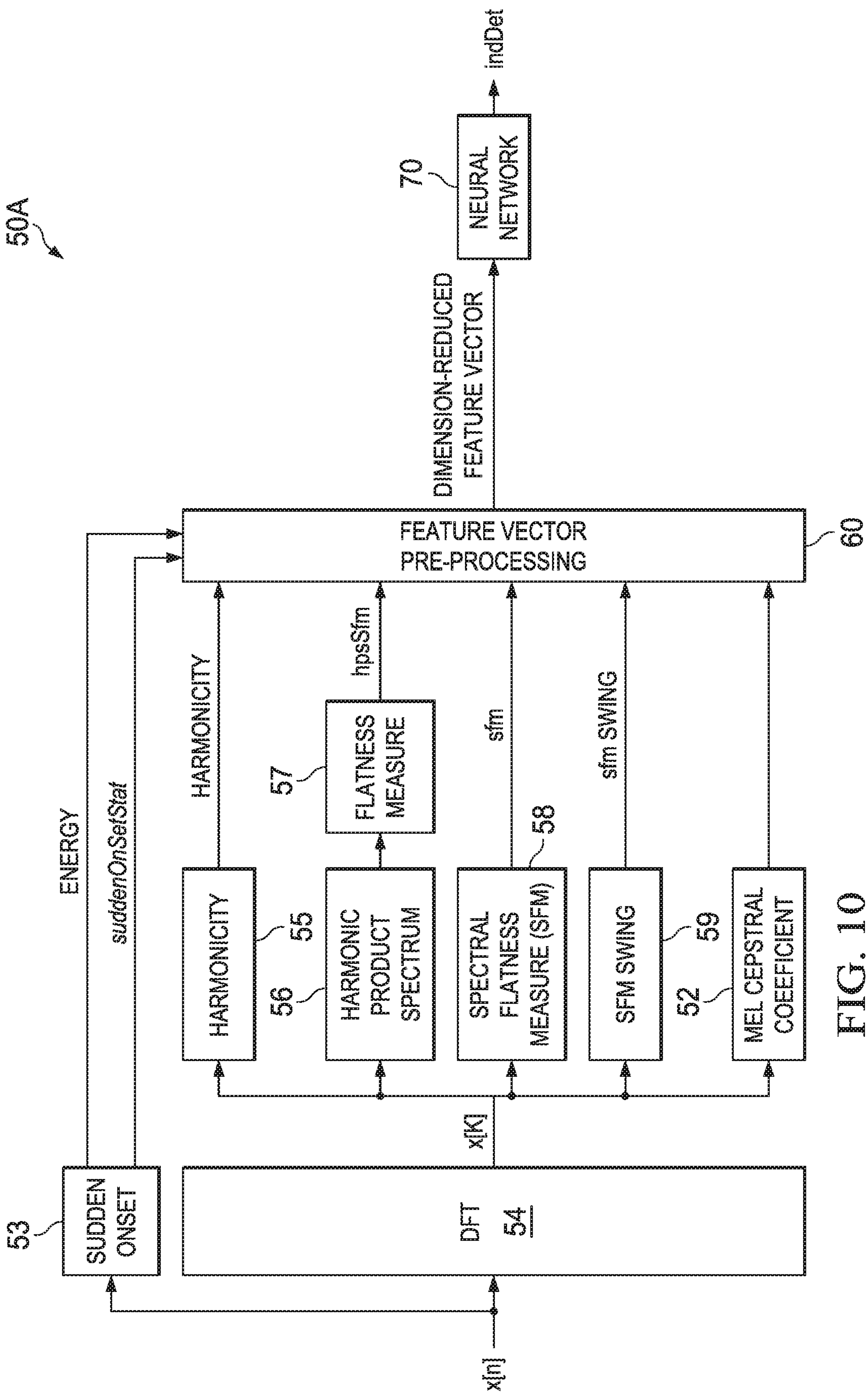

50A
x[n]
DFT 54
53
SUDDEN ONSET
ENERGY
suddenOnSetStat
x[K]
HARMONICITY
55
HARMONIC PRODUCT SPECTRUM
56
SPECTRAL FLATNESS MEASURE (SFM)
58
SFM SWING
59
MEL CEPSTRAL COEFFICIENT
52
FLATNESS MEASURE
57
HARMONICITY
hpsSfm
sfm
sfm SWING
FEATURE VECTOR PRE-PROCESSING
60
DIMENSION-REDUCED FEATURE VECTOR
NEURAL NETWORK
70
indDet

FIG. 10

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2017/055887

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L21/0216 G10L25/30
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GB 2 456 296 A (SEPEHR HAMID [GB]; NOORALAHIYAN AMIR [GB]) 15 July 2009 (2009-07-15) page 6 - page 7; figures 1, 2, 8 page 9 - page 13 page 17 - page 20; claim 13 -----	1-18
X	POTAMITIS I ET AL: "Impulsive noise suppression using neural networks", ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2000. ICASSP '00. PROCEEDING S. 2000 IEEE INTERNATIONAL CONFERENCE ON 5-9 JUNE 2000, PISCATAWAY, NJ, USA,IEEE, vol. 3, 5 June 2000 (2000-06-05), pages 1871-1874, XP010507728, ISBN: 978-0-7803-6293-2 page 1871 - page 1873 -----	1-18

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 January 2018	02/02/2018

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Belardinelli, Carlo

1

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	MARCO RUHLAND ET AL: "Reduction of gaussian, supergaussian, and impulsive noise by interpolation of the binary mask residual", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE, USA, vol. 23, no. 10, 1 October 2015 (2015-10-01), pages 1680-1691, XP058072921, ISSN: 2329-9290, DOI: 10.1109/TASLP.2015.2444664 abstract paragraph [000I] - paragraph [0III] -----	1-18
A	Andrzej Czyzewski: "Learning Algorithms for Audio Signal Enhancement, Part 1: Neural Network Implementation for the Removalof Impulse Distortions*", JAES Volume 45 Issue, 31 October 1997 (1997-10-31), pages 815-831, XP055442324, Retrieved from the Internet: URL:http://www.aes.org/e-lib/inst/download.cfm/7839.pdf?ID=7839 [retrieved on 2018-01-18] the whole document -----	1-18

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2017/055887

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
GB 2456296 A	15-07-2009	NONE	