



(51) International Patent Classification:
G10L 15/06 (2013.01)

(21) International Application Number:
PCT/AU2012/001331

(22) International Filing Date:
31 October 2012 (31.10.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/553,746 31 October 2011 (31.10.2011) US

(71) Applicant (for all designated States except US): **THE UNIVERSITY OF MELBOURNE** [AU/AU]; Grattan Street, Parkville, Victoria 3052 (AU).

(72) Inventors; and

(71) Applicants (for US only): **MCLACHLAN, Neil, Maxwell** [AU/AU]; c/- UoM Commercial Limited, 205-211 Grattan Street, Carlton, Victoria 3053 (AU). **DEGHANI, Arvin** [AU/AU]; c/- UoM Commercial Limited, 205-211 Grattan Street, Carlton, Victoria 3053 (AU).

(74) Agent: **DAVIES COLLISON CAVE**; 1 Nicholson Street, Melbourne, Victoria 3000 (AU).

(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, IS, JP, KE, KG, KM, 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, 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, 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,

[Continued on next page]

(54) Title: A SIGNAL PROCESS, A SIGNAL RECOGNITION PROCESS AND A SIGNAL RECOGNITION SYSTEM

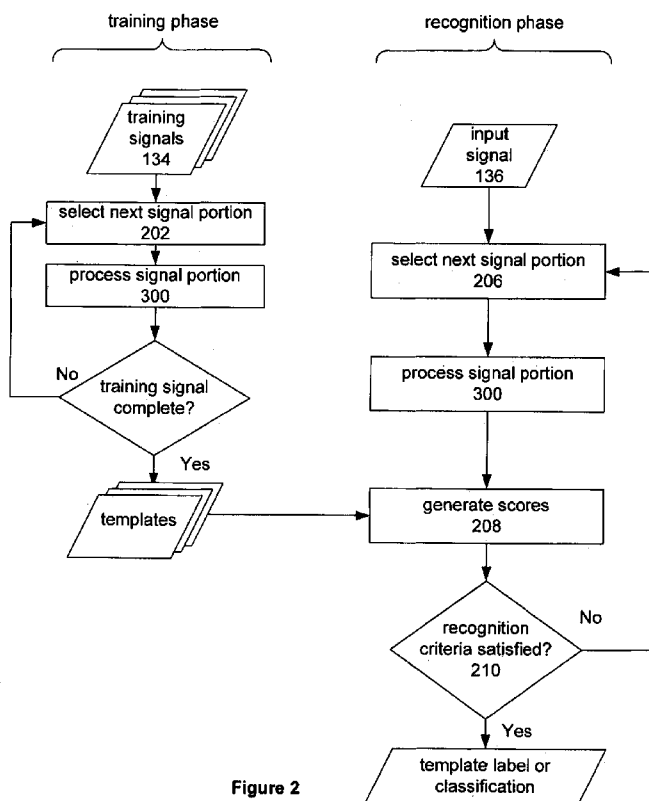


Figure 2

(57) Abstract: A signal recognition process, including: receiving signal data representing a signal; filtering the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data; setting signal amplitudes exceeding a saturation threshold to a saturation value representing reinforcement; and applying lateral inhibition across each of the one or more other dimensions to generate, for each said other dimension, inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension.



TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, — *with amended claims (Art. 19(1))*
ML, MR, NE, SN, TD, TG).

Published:

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

- 1 -

A SIGNAL PROCESS, A SIGNAL RECOGNITION PROCESS AND A SIGNAL RECOGNITION SYSTEM

TECHNICAL FIELD

The present invention relates to a signal process, a signal recognition process and a signal
5 recognition system.

BACKGROUND

There are many situations in which it is desired to recognise or identify a signal as being a
member of a known class or corresponding to a known type of signal. For example, there
10 may be a need to recognise a sound as being a component of human speech, or as a
particular spoken word, or as being a particular type of musical sound (*e.g.*, a major
chord). Although computer-implemented signal processing methods for recognising input
signals do exist, they have limited capabilities and performance.

15 It is desired to provide a signal recognition process, a signal process, and a signal
recognition system that alleviate one or more difficulties of the prior art, or that at least
provide a useful alternative.

SUMMARY

In accordance with some embodiments of the present invention, there is provided a signal
recognition process, including:

- receiving signal data representing a signal;
- 20 filtering the signal data to generate filtered data representing signal
amplitudes as a function of time and one or more other dimensions represented by
the signal data;
- setting signal amplitudes exceeding a saturation threshold to a saturation
value representing reinforcement; and
- 25 applying lateral inhibition across each of the one or more other dimensions
to generate, for each said other dimension, inhibitive signal amplitude values at

- 2 -

values of said dimension flanking dominant ones of the signal amplitudes along said dimension.

In some embodiments, the signal recognition process includes applying temporal inhibition to the signal amplitudes to produce inhibitive signal amplitude values immediately following offsets of the saturated signal amplitudes.

In accordance with some embodiments of the present invention, there is provided a signal recognition process, including:

- 5 receiving signal data representing a signal;
- filtering the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data;
- applying lateral inhibition across each of the one or more other dimensions
- 10 to generate inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension; and
- applying temporal inhibition to the signal amplitudes to produce inhibitive signal amplitude values immediately following offsets of the saturated signal amplitudes.

15

In accordance with some embodiments of the present invention, there is provided a signal recognition process, including:

receiving training signal data representing one or more training signals and processing the training signal data to generate signal recognition templates using the process of any one of the above processes;

receiving input signal data representing an input signal to be recognised and processing the input signal data to generate processed input signal data using the process of any one of the above processes;

for each of the signal recognition templates, generating a corresponding recognition score quantifying correspondence between the processed input signal data and the signal recognition template.

- 3 -

In some embodiments, the signal recognition process includes selecting, on the basis of the generated scores, at least one of the signal recognition templates as corresponding to the input signal.

In some embodiments, the signal recognition process includes determining the saturation threshold such that a specified proportion of the signal amplitudes exceed the saturation threshold.

In some embodiments, the step of setting signal amplitudes includes, for each of a plurality of saturation thresholds, generating a corresponding set of recognition templates in which signal amplitudes exceeding the corresponding saturation threshold are set to a saturation value representing reinforcement.

5

In some embodiments, the signal recognition process includes:

generating, for each of a plurality of time windows of each of said templates, a corresponding decision value based on the corresponding amplitude values of the template;

10

for each of a plurality of time windows following a detected signal onset, generating dot products of the corresponding positive amplitude values of the processed input signal data and the corresponding amplitude values of respective ones of the signal recognition templates; and

15

for each of the plurality of time windows following the detected signal onset, comparing a maximum one of the generated dot products with the corresponding decision value for the time window, and determining whether the corresponding signal recognition template is a match to the input signal for the time window based on said comparing; and

20

selecting at least one of said signal recognition templates as being recognised based on the number of matches of the at least one signal recognition template to the input signal.

- 4 -

In some embodiments, the signal recognition process includes reducing the number of said signal recognition templates by combining similar ones of said signal recognition templates identified by generating scores quantifying correspondence between at least some of said signal recognition templates.

In some embodiments, a plurality of said signal recognition templates are generated for successive temporal portions of each training signal, and each recognition score is generated from a template for a corresponding temporal portion of a training signal and processed input signal data for a corresponding temporal portion of the input signal.

In some embodiments, the received input signal represents a combination of a first signal corresponding to one of the training signals and at least one second signal overlapping with the first signal, the selected signal recognition template corresponds to a first temporal portion of the first signal, and the process further includes determining predicted first signal data on the basis of a further signal recognition template corresponding to a second temporal portion of the first signal subsequent to the first portion of the first signal, and using the predicted first signal data to improve recognition of the at least one second signal.

In some embodiments, the signal recognition process includes generating one or more background templates from unrecognised temporal portions of the input signal, and using the generated background templates to improve the recognition of input signal components in subsequent temporal portions of the input signal.

In some embodiments, a plurality of said signal recognition templates are generated for successive temporal portions of each training signal, and each recognition score is generated from a template for a corresponding temporal portion of a training signal and processed input signal data for a corresponding temporal portion of the input signal.

- 5 -

In accordance with some embodiments of the present invention, there is provided a signal process, including:

(i) for each of a plurality of training signals, generating a set of signal templates representing successive temporal portions of the training signal;

(ii) processing successive temporal portions of an input signal to generate respective processed input signal portion data;

(iii) selecting a subset of the signal templates representing selected temporal portions of each training signal and processed input signal portion data representing a corresponding selected temporal portion of the input signal;

(iv) for each said training signal, processing the corresponding signal template and the selected processed input signal portion data to generate a corresponding score representing correspondence between the selected temporal portion of the training signal and the selected temporal portion of the input signal;

(v) selecting a further subset of the signal templates representing a subsequent temporal portion of each training signal and processed input signal portion data representing a corresponding further temporal portion of the input signal; and

(vi) repeating step (iv) to generate further scores for the further temporal portions.

In some embodiments, in step (v) only signal templates from sets of templates whose scores generated at step (iv) exceeded a threshold value are selected.

In some embodiments, step (vi) includes repeating step (iv) until the generated scores satisfy one or more predetermined criteria.

In some embodiments, the process of processing the corresponding signal template and the selected processed input signal portion data to generate a corresponding score representing correspondence between the selected temporal portion of the training signal and the selected temporal portion of the input signal is substantially a real-time process.

In some embodiments, the one or more other dimensions include frequency. In some embodiments, the one or more other dimensions include one or more spatial dimensions. In

- 6 -

some embodiments, the signal includes an audio signal. In some embodiments, the signal includes a video signal.

In some embodiments, the process is configured to recognise sounds. In some embodiments, the process includes the process is configured to recognise human speech.

In accordance with some embodiments of the present invention, there is provided a computer-readable storage medium having stored thereon processor-executable instructions that, when executed by a processor, cause the processor to execute the process of any one of the above processes.

In accordance with some embodiments of the present invention, there is provided a signal recognition system configured to execute any one of the above processes.

In accordance with some embodiments of the present invention, there is provided a signal recognition system, including:

a signal processing component configured to:

(i) receive signal data representing a signal;

(ii) filter the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data;

5 (iii) set signal amplitudes exceeding a saturation threshold to a saturation value representing reinforcement; and

(iv) apply lateral inhibition across each of the one or more other dimensions to generate, for each said other dimension, inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension;

a training component configured to receive training signal data representing one or more training signals and to cause the signal processing component to process the training signal data to generate signal recognition templates; and

a signal recognition component configured to:

- 7 -

(a) receive input signal data representing an input signal to be recognised and to cause the signal processing component to process the input signal data to generate processed input signal data; and

(b) for each of the signal recognition templates, to generate a corresponding recognition score quantifying correspondence between the processed input signal data and the signal recognition template.

BRIEF DESCRIPTION OF THE DRAWINGS

Some embodiments of the present invention are hereinafter described, by way of example only, with reference to the accompanying drawings, wherein:

5 Figure 1 is a schematic block diagram of a signal recognition system in accordance with some embodiments of the present invention;

Figure 2 is a flow diagram of a signal recognition process executed by the signal processing system;

Figure 3 is a flow diagram of a signal process of the signal recognition process;

10 Figure 4 is a schematic illustration of lateral inhibition across frequencies for a particular time slice of an input or training signal processed by the signal process of Figure 3;

Figure 5 is a graph of Gammatone filterbank output for a 7-semitone chord of seven equal-amplitude harmonic complexes;

15 Figure 6 is a graph of signal amplitude as a function of frequency generated from an input signal: (i) as output by the Gammatone filterbank (dot-dash line), (ii) after lateral inhibition (solid line), and (iii) after saturation (dotted line);

20 Figures 7 (a), (b), and (c) are surface plots illustrating the dependence of recognition accuracy on the two recognition parameters inhibition width and inhibition strength for Gammatone filterbanks with 22, 50 and 300 channels, respectively;

Figures 8 (a), (b), (c) and (d) are surface plots illustrating the dependence of recognition accuracy on the two recognition parameters inhibition width and saturation

- 8 -

density for (a) 50 filterbank channels and $I_s=0.055$, (b) 300 filterbank channels and $I_s=0.055$, (c) 50 filterbank channels and $I_s=0.075$, and (d) 300 filterbank channels and $I_s=0.075$;

Figure 9 is a graph of recognition accuracy as a function of the number of
5 filterbank channels (see text for details);

Figure 10 is a graph of the amplitude of a frequency channel as a function of time generated from an input signal: (i) as output by the Gammatone filterbank (dot-dash line), (ii) after saturation (solid line), and (iii) after temporal inhibition (dotted line);

Figures 11 to 14 are graphs of template activation as a function of time for 0, 5, 10
10 and 20 Hz AM templates, respectively, and for a harmonic complex at 172 Hz with 0, 5, 10 and 20 Hz AM presentations of a harmonic complex stimuli at 175 Hz;

Figure 15 is a graph of template activation as a function of input signal AM frequency for 175 Hz harmonic complex templates with 0, 5, 10, 15 and 20 Hz AM and input signals with varying AM rates 150 ms after onset;

15 Figure 16 is a flow diagram of an alternative signal process to that of Figure 3;

Figure 17 is a set of four plots representing respective templates (as amplitude as a function of filter channel and time) for the spoken phoneme "ma" generated for intensity thresholds levels of (a) 1, (b) 5, (c) 9 and (d) 12;

Figure 18 is a plot of a merged template for saturation level 1 for all the phonemes
20 used in the second Example; and

Figure 19 is a graph of recognition accuracy as a function of input signal loudness (see text for details); and

Figure 20 is a graph of recognition accuracy as a function of background noise (see text for details).

DETAILED DESCRIPTION

Described herein are signal recognition processes that process input signals in order to recognise or classify at least a portion of an input signal as being an instance or example of a particular class or type of signal as represented by one or more signal templates.

- 5 Although the signal processes are generally described herein in the context of processing audio signals and even more particularly in the context of recognising musical sounds such as chords, the invention is not limited to such applications and may have broad application to other fields, including environmental sound recognition for noise, defence, and security monitoring applications, music transcription and retrieval, and automated speech
- 10 recognition. More generally, these processes can be applied to any signal with time varying amplitudes that can be decomposed into one or more other dimensions, wherein the lateral inhibition processes described herein in relation to the frequency domain for audio signals are applied to all dimensions other than amplitude and time. For example, the processes can be applied to video signals, and the other dimensions can include one or two spatial
- 15 dimensions represented by the video signal, and optionally also the frequency dimension. Other dimensions, signal types and applications of the described processes will be apparent to those skilled in the art in light of this disclosure.

- Broadly, the signal recognition processes described herein include: (i) a training process or
- 20 phase that processes training input signals to generate corresponding sets of signal templates representing different classes, categories or labels of signal, and (ii) a recognition or classification process or phase that processes a subsequently received input signal in order to 'recognise' that signal as corresponding to one or more of the templates, and hence to 'recognise' it as an instance of at least one corresponding class, category or
- 25 label of signal.

- When applied to audio signals representing sounds, the described signal recognition processes mimic to some extent the neurobiology of the human auditory system, which has evolved to rapidly recognise sounds that may represent, for example, sounds of imminent
- 30 danger, or human speech. Thus an input signal may be recognised as, for example, a major third chord, or human speech or a particular vowel sound, or the sound of a submarine

- 10 -

propeller, or the sound of a missile launch, or the sound of a failing mechanical bearing, or essentially any other type of sound.

In the described embodiments, template matching is applied to the amplitudes of a set of
5 bandpass filters of up to 300 ms of sound. Each template is generated from sequential spectral 'time slices' of rectified filter outputs integrated over temporal windows of up to about 10 ms. Where the signal recognition process is applied to recognition of musical chords, the bandpass filter resolution is selected to be sufficient to segregate individual lower order harmonics in signals, but not to be so fine that it generates a proliferation of
10 templates with similar frequency information. Filterbank properties based on the human auditory system can be used, as they have evolved for this purpose. A set of templates is generated from training signals representing sounds that are exemplars of one or more sound source labels/identities/classes/categories (where these terms are used interchangeably herein). A single identity can be associated with multiple templates that
15 vary along an acoustic dimension such as fundamental frequency, thereby making that identity invariant along that dimension. The duration and spectral and temporal resolution of the templates can be selected according to the properties of the sound to be recognized.

The generated templates have spectro-temporal regions of excitation/reinforcement where
20 the sound amplitudes are high. The accuracy of the template matching is greatly enhanced by including bands of inhibition surrounding regions of excitation in both the spectral and temporal dimensions. In this embodiment, this is achieved in the spectral dimension by integrating the filter outputs over a running window of desired width of the lateral inhibition band (a user defined input variable). The integrand is then multiplied by a
25 weighting factor and subtracted from the centre channel of each integration window, thereby creating bands of inhibition on either side of regions of excitation. In some embodiments, a similar approach is applied to the temporal dimension, except that the weighted integrand is subtracted from the last time point of the integration window. This means that the level of temporal inhibition in the template is proportional to the recent
30 levels of excitation in each channel.

- 11 -

Recognition is insensitive to overall loudness and is often robust to variation in the amplitude of spectral components. In some embodiments, this is achieved by saturating the signal amplitudes that exceed a dynamically determined threshold value. In the first embodiment described below, saturation is applied after spectral lateral inhibition and prior
5 to temporal inhibition, although in other embodiments (including the second embodiment described below), it is done first, which has been found to be more effective. In some embodiments, the saturation threshold is dynamically determined so that a specified proportion of channels will be driven to saturation. This ensures that the sum of excitation is similar for all templates, and so a greater amplitude of spectro-temporal components in
10 one template do not increase the likelihood of its recognition compared to similar templates.

During classification, the excitatory component of each temporal portion or slice of an input signal (after saturation and inhibition) is applied to the first temporal slice of each
15 template in the array. In some embodiments, only if the activation of a template exceeds a user-defined threshold value, then the second temporal slice of the template is applied to next temporal slice of the signal. The signal continues to be compared to successive temporal portions of a particular template as long as the activation of that template remains above a user-defined threshold, which may vary according to the desired sensitivity of the
20 recognition process to target sounds. While many templates can be activated by the onset information, sequentially fewer templates will remain activated as temporal information becomes available, thereby reducing the computational load.

In other embodiments, all the temporal slices are computed for all templates and the
25 maximally activated template by a portion of the test signal is selected as representative of the test signal. Furthermore, contextual information (such as the time of day, for example) can be used to alter the likelihood that certain templates will be activated by applying appropriate weights to templates of identities with more or less probability of occurring at that time (or other context(s)).

- 12 -

Each of the templates generated by the training process can be considered as a multi-dimensional array of weights representing varying degrees of spectro-temporal activation (where the weight values are positive) or inhibition (where the weight values are negative), so that the presence of frequency and/or temporal attributes in an input signal corresponding to those (positively weighted) in a template positively weight or activate that template towards recognition, whereas the presence of frequency and/or temporal attributes in the input signal that are absent from the training signal are negatively weighted in the template to inhibit that template away from recognition. In this way, an overall weight or score representing the cumulative degree of activation/inhibition is generated for each template.

In some embodiments, a vector of different saturation threshold values is defined, and templates generated using each of the values of the saturation threshold vector are compared to signal representations generated using all the values of the saturation threshold vector. The largest score generated for each time point across all values of the saturation threshold is then compared to the activation threshold for that template, thereby providing high tolerance to variation in the signal amplitude.

In neurological terms, the templates are said to be 'activated', by analogy with neural activation. As the input signal evolves with time, the scores are progressively updated and thus the specificity and certainty of recognition can increase over time. The certainty of recognition can be quantified as the ratio of activation between the strongest and second strongest activated templates. Recognition processing can stop when the certainty of recognition reaches an acceptable level in a given context, or when the input signal ends. Optionally, in order to reduce the computational load, any templates whose scores are below a cut-off threshold value can be eliminated from further processing of the input signal, albeit at the risk of reduced recognition performance for some inputs (e.g., phonemes). As templates that were initially activated (i.e., whose scores were above the cut-off threshold value) are inhibited by off-frequency and/or off-time components of the input signal that arrive in subsequent time steps their activation may drop below the cut-off threshold value. In any case, the input signal is then considered to be 'recognised' as the

- 13 -

label(s) or classification(s) of the remaining template(s) or subset thereof, depending on the respective activation values.

As described below, the training component of the process generates the signal templates
5 by: (i) applying lateral inhibition to positively reinforce dominant frequencies while
inhibiting the presence of other close frequencies with lesser amplitudes (referred to herein
as 'off-frequencies'), (ii) applying rate saturation to make the signal recognition
substantially independent of absolute signal amplitudes, and (iii) applying temporal
10 inhibition to positively reinforce recognition of a template when signal onsets coincide
with the onset of positive information in the template and to inhibit recognition of a
template where the input signal has substantial positive amplitudes at times where the
exemplary signal from which the template was generated does not and that follow periods
of high signal amplitudes.

15 In the described embodiments, the signal recognition processes are implemented as one or
more software modules executed by a standard computer system such as an Intel IA-32
based personal computer system, as shown in Figure 1. However, it will be apparent to
those skilled in the art that at least parts of the signal processes could alternatively be
implemented in part or entirely in the form of one or more dedicated hardware
20 components, such as application-specific integrated circuits (ASICs) and/or field
programmable gate arrays (FPGAs), for example. Moreover, the signal recognition
processes could be implemented as software for low power computing and/or digital signal
processing devices, including portable devices such as 'smart-phones', hearing aids and the
like.

25

As shown in Figure 1, a signal recognition processing system 100 executes a signal
recognition process, as shown in Figure 2, which is implemented as one or more software
modules 102 stored on non-volatile (e.g., hard disk, solid-state drive, or flash memory)
storage 104 associated with a standard computer system. The system 100 includes
30 standard computer components, including random access memory (RAM) 106, at least one
processor 108, and external interfaces 110, 112, 114, 115, all interconnected by a bus 116.

- 14 -

The external interfaces include universal serial bus (USB) interfaces 110, at least one of which is connected to a keyboard 118 and a pointing device such as a mouse 119, a network interface connector (NIC) 112 which can be used to connect the system 100 to a communications network such as the Internet 120, a display adapter 114, which is
5 connected to a display device such as an LCD panel display 122, and a sound card 115, which is connected to a microphone 124 and optionally a speaker 126. The system 100 also includes a number of standard software modules 128 to 132, including an operating system 128 such as Linux or Microsoft Windows, and the Matlab software package 130 and the Auditory Toolbox 132 for Matlab. An example Matlab code listing implementing
10 the signal process is included as an Appendix to this specification.

As shown in the flow diagram of Figure 2, the signal recognition process begins by receiving or accessing training signals 134 in the form of digitised training signal data at step 202. Typically, but not necessarily, each training signal 134 is received or accessed in
15 the standard form of a stream of 16-bit digitised audio samples acquired at a sampling rate of 16,000 Hz and encoded in standard linear pulse code modulation (LPCM) format, and may be stored in a 'WAV' container. However, it will be apparent to those skilled in the art that a wide variety of other amplitude resolutions, sampling rates, audio codecs and formats may be alternatively used. The training signals 134 and/or input signals 136 may
20 be generated on the system 100 by a user (e.g., via the sound card 115 and microphone 124) and may be either stored in encoded form for asynchronous or off-line processing at a later time, or alternatively processed in real-time during receipt or generation. Alternatively, some or all of the training signals 134 and/or the input signals 136 may be received from the network 120 via the NIC 112. Some or all of the training signals 134
25 and/or the input signals 136 may be stored as one or more encoded data files on the non-volatile storage 104 of the signal recognition system 100.

For the purpose of providing indicative values for signal and process parameters, the signal recognition processes are described herein in the context of musical sound recognition,
30 wherein the input signal data 134 represents a digitally sampled audio signal representing musical sounds such as pure tones and multi-tone chords. However, as already indicated

- 15 -

above, the signal recognition processes have broad application to a wide variety of different recognition tasks, and the input signal could alternatively represent other types of sound and/or video signals, including human speech, for example, or indeed non-audio-visual signals.

5

The Training Phase

As shown in Figure 2, the training phase of the signal recognition process operates on successive temporal portions or 'time slices' of fixed (but user-configurable) duration (default value 10 ms) of each training signal 134 and generates templates from those
10 temporal portions, so that each template consists of a temporal sequence of three-dimensional spectral time slices (representing signal amplitude as a function of frequency and time) up to a configurable maximum signal duration. By default, the maximum signal duration is 300 ms, but this duration can be changed by the user as desired (subject to memory constraints). After selecting the next time slice 204 of a training signal 134 at step
15 202, the selected time slice or portion 302 of the training signal 134 is then processed by a signal process 300, as shown in Figure 3.

Where the application (e.g., musical sound recognition) models the human auditory system, a multi-channel Gammatone filterbank is applied at step 304 to divide the signal
20 portion 302 into overlapping frequency bands or channels, in this case, between 50 Hz and 5,000 Hz. The frequencies and sampling rates in the described embodiments have been selected for recognition of musical sounds. However, it will be apparent to those skilled in the art that different values for these process parameters can be used in other embodiments, depending largely on the nature of the frequency spectrum of the sounds that are to be
25 recognised.

The temporal signal portion as processed up to this point can be represented as a two-dimensional array $y_g(f, t)$ and thought of as a three-dimensional cuboid or rectangular prism of data with dimensions representing signal amplitude (as a function of) time, and
30 frequency. The following steps of the signal process 300 operate individually on each of these three dimensions.

- 16 -

At step 306, the signal in each channel is half-wave rectified (*i.e.*, any negative signal amplitudes are set to zero) to approximate hair-cell transduction in the human auditory system, as follows:

$$y_r = \begin{cases} y_g, & y_g \geq 0 \\ 0, & \text{else} \end{cases}$$

and at step 308, the half-wave rectified signal y_r is integrated over successive 10 ms integration windows to collapse the time dimension and thereby provide a single spectral 'time slice' $y_w(n)$ representing amplitude as a function of frequency channel, as follows:

$$j = i T_w$$

$$y_w(n, j) = \sum_{k=0}^{T_w} y_r(n, j + k), \quad 1 \leq i \leq L - T_w$$

where n represents the filter (*i.e.*, frequency band) channel, i represents the data sample, T_w represents the length of the integration window and L represents the length of the signal. A more accurate hair-cell model is unnecessary, as integration over the 10 ms window incorporates the temporal dynamics of both slow and fast refractory auditory nerves. The signal integration step 308 may not be required in some applications, or the integration window may of course differ, depending on the sample rate(s) of the training and input signals 134, 136, the temporal nature of the signals 134, 136, and the recognition requirements.

20

Lateral Inhibition

At step 312, lateral inhibition is applied across the frequency (*i.e.*, filterbank channel) dimension to enhance template recognition where dominant frequencies in a training signal are present in an input signal and to inhibit template recognition where off-frequency signal components are present in the input signal. Specifically, the signal data across filterbanks is first processed according to:

$$y_l(n) = y_w(n) - I_s M(n), \quad 0 \leq n \leq N, \quad 0 \leq I_s, \quad I_w > 0$$

- 17 -

$$M(n) = \begin{cases} 2/I_w \sum_{k=1}^{I_w} y_w(n+k), & n \leq I_w \\ 1/I_w \sum_{k=1}^{I_w} y_w(n+k) + y_w(n-k), & 1 + I_w \leq n \leq N - 1 - I_w \\ 2/I_w \sum_{k=1}^{I_w} y_w(n-k), & n \geq N - I_w \end{cases}$$

5

where I_s and I_w are adjustable parameters representing lateral inhibition strength and lateral inhibition width, respectively. For example, where the process is applied as described above to recognise musical sounds with the process parameters described above, the lateral inhibition width I_w can be set to 7 filter channels, and the lateral inhibition strength I_s can be set to 0.055. That is, each amplitude value in each filter channel is adjusted by subtracting 5.5% of the sum of the current amplitude value with the corresponding amplitude values from the adjacent 7 filter channels on either side of the current filter channel value. Additionally, all of the resulting amplitude values in the template are then reduced by a fixed (but user-configurable) proportion (10% by default) of the modified maximum peak value in the template.

The resulting amplitude values have a characteristic 'Mexican hat'-like shape across the frequency channels, with a positive valued peak at the dominant frequency between a pair of inhibitive (*i.e.*, negative-valued) side lobes, as shown schematically in Figure 4. With the values described above, the inhibitive side lobes were located at about $\pm 8\%$ of the corresponding peak frequency, with amplitudes of about -2% to -10% of the peak amplitude. In practice, however, these values can be selected (*e.g.*, by trial and error) to provide good recognition of particular sounds, depending on the application requirements.

25 *Saturation*

In general, it is desirable that recognition performance be independent of sound pressure levels. For example, a musical sound or a spoken vowel sounds should be recognised independently of the loudness of those sounds. Consequently, the template selection process should be substantially independent of the amplitudes of the training signals 134. Otherwise, the mean amplitudes of training signals 134 could artificially bias the template

selection process so that, for example, templates generated from training signals with high amplitudes could be more likely to be selected than templates generated from similar training signals but having lower amplitudes.

- 5 For example, Figure 5 shows the output of a Gammatone filter bank for a 7-semitone chord (a frequency ratio of 1.5). The third and sixth harmonics of the lower frequency complex are at the same frequency as the second and fourth harmonics (respectively) of the higher frequency complex, resulting in two peaks of substantially higher amplitude than the other peaks. The mean amplitudes of 7-semitone chords are therefore lower than for chords that
10 do not contain pairs of closely tuned harmonics for input signals that are initially normalized by their peak amplitudes.

To ensure that the mean amplitudes of all templates are similar, in some embodiments the signal process 300 dynamically determines an amplitude threshold such that a user-
15 configurable proportion of all filter channels exceed that amplitude threshold. In the training phase, the determined amplitude threshold remains constant for all filter channels and time slices of each training signal. In the recognition phase (as described below), the amplitude threshold can be determined in the same way, or alternatively (and more efficiently computationally) where appropriate, the amplitude threshold can be determined
20 in relation to recent events so that background sounds are not saturated and the average input signal is saturated to the specified extent. The saturation threshold is usually at around 10% of the maximum filter channel amplitude, and thus allows for a wide variation in input signal amplitude.

- 25 Once this amplitude threshold value has been determined, at step 310 all values exceeding the threshold values are set to a maximum or saturation value of 1.0, as follows:

$$y_s(n) = \begin{cases} y_t(n), & \text{if } y_t(n) \leq S, \\ 1, & \text{otherwise} \end{cases}$$

- 19 -

For the example application to musical recognition described herein, the configurable proportion of all filter channels (referred to herein as the saturation density S_d) was set to 0.24, which resulted in a saturation threshold value S_t that was typically about 4% of the peak amplitude.

5

Temporal Inhibition

At step 314, the signal process 300 generates temporal inhibition fields for each temporal portion 302 of the signal by summing the amplitudes of recent time steps in each frequency channel using a running integration window, scaling the resulting sum, and then
10 subtracting the scaled portion from the current amplitude value, as follows:

$$\hat{y}_s(j) = y_s(j) - \hat{I}_s \hat{M}(j), \quad 1 \leq j \leq N_t - 1, 0 \leq \hat{I}_s \leq 1, \hat{I}_w > 0$$

$$\hat{M}(j) = \begin{cases} \sum_{k=0}^{\hat{I}_w} y_s(j-k), & \hat{I}_w \leq j \leq N_t - 1 \\ \sum_{k=0}^j y_s(k) \cdot \frac{\hat{I}_w + 1}{j + 1}, & j < \hat{I}_w \end{cases}$$

As will be seen in the Example below, this creates sharp temporal transitions that boost amplitude values at the onset of positive portions of the signal and generate negative
15 amplitude values immediately after offsets in the resulting processed signal portions 316.

Returning to the flow diagram of Figure 2, the signal recognition process loops back to step 202 to retrieve the next temporal portion of the training signal 134 until all the temporal portions of the training signal 134 have been processed. The processed temporal
20 portions of the training signal 134 are stored on the signal recognition system 100 and constitute signal recognition templates 138 for the training signal 134. Thus each template can be considered to consist of a series of spectral slices, each spectral slice being an array of amplitudes as a function of frequency channel for that time slice (or 10 ms portion of the corresponding signal). These steps are then repeated for additional ones of the training
25 signals 134 to generate additional signal recognition templates 138. A user of the system 100 provides labels for subsets of one or more of the signal recognition templates 138. This completes the training phase.

The Recognition Phase

- The recognition phase of the signal recognition process begins when an input signal 136 to be recognised is received. The first temporal portion of the input signal 136 is selected at step 206 as described above for the training signal 134, and the selected temporal portion is identically processed by the signal process 300 as described above to produce, initially, a single spectral slice for the first time slice (in the described embodiment, representing a temporal duration of 10 ms).
- At step 208, numeric scores quantifying the degree of correspondence of the initial spectral slice of the input signal 136 to the initial spectral slice of each of the respective templates 138 are generated. Each score is generated as the sum of the products of only the positive amplitudes of the spectral slice of the input signal 136 with the corresponding positive or negative amplitudes of the corresponding spectral slice of the corresponding template.
- At step 210, the resulting scores are compared to assess the certainty of recognition, and if this meets user-specified requirements (as described below), then the input signal 136 is deemed to be successfully recognised, and the corresponding label associated with the highest scoring template(s) is provided as output. Otherwise, the process loops back to select the next portion of the input signal 136 at step 206. In the case of the first spectral slice of the input signal, there will in general be many templates whose first spectral slice 'matches' that of the input signal. Consequently, in some embodiments the process will typically loop back to select and process the second spectral time slice of the input signal 136 and to compare that second slice with the second slice of each template whose initial score for the previous slice was sufficiently high. That is, in some embodiments other templates with lower scores will not in general have their second spectral slices compared to the second spectral slice of the input signal. However, until a positive recognition has occurred, the first spectral slices of all templates will continue to be compared to the second spectral slice (and subsequent spectral slices) of the input signal 136. In other embodiments, all templates continue to be compared with the input signal; although this

- 21 -

increases the computational load, it can improve the recognition accuracy for some input signals (*e.g.*, phonemes).

In either case, this general process continues until one or more user-specified recognition
5 criteria are satisfied, typically being that the certainty of recognition exceeds a specified threshold value, or more specifically, that (i) at least one of the scores exceeds a first user-specified threshold value, and also (ii) the ratio of the two highest scores exceeds a second user-specified threshold value.

10 Thus the very first spectral slice of each and every template is processed (in parallel) against the current spectral time slice of the input signal 136, searching for matches. Optionally, in order to reduce the computational load, the system 100 can be configured so that only those templates whose scores for the previous one or more spectral slices exceed a configurable threshold value have their second or later spectral slices processed against
15 the current spectral time slice of the input signal. However, the resulting reduction in computational load may be at the cost of reduced recognition performance for some inputs (*e.g.*, phonemes). This arrangement greatly reduces the computational overhead of processing, and allows the signal recognition process to be implemented on relatively low power computing devices, including portable devices such as smart phones and the like.

20

As described above, the certainty of recognition can be quantified as the ratio of activation between the strongest and second strongest activated templates (*i.e.*, as the ratio of the two highest scores). Recognition processing can stop when the certainty of recognition reaches an acceptable level as defined by the user. The input signal 136 is then considered to be
25 'recognised' as the label(s) or classification(s) associated with the highest scoring template(s). A single label or identity can be associated with multiple templates that vary along an acoustic dimension such as fundamental frequency, thereby making that label invariant along that dimension. For speech recognition, this feature can be used to associate a single label (*e.g.*, an identified vowel or word, for example) with multiple
30 exemplary training signals presenting that vowel or word spoken by persons with different accents and/or speaking rates and/or frequencies, for example.

- 22 -

The ability of the signal recognition process to dynamically process input signals in real-time also allows it to dynamically track multiple signal components (corresponding to respective labels/categories: for example, different sounds) over time as the activation of templates changes over time. For example, where the signal recognition process is used to process input signals representing sounds, changes in those sounds over time (*e.g.*, as different speakers begin and stop speaking, and/or different musical instruments begin or stop being played) can be tracked.

In general, each time the first spectral slice of a template has been activated, the sequence of comparing subsequent spectral slices with the incoming signal is initiated independently of the processing of other templates. However, when one sound occurs before another, and strongly activates a template, then the template information for subsequent time slices can be used to subtract the predicted spectro-temporal information associated with the training signal corresponding to the activated template from the spectral slices of the input signal in order to enhance the sensitivity of recognising the other sound.

Similarly, input signals that do not match any templates can be automatically assigned to a background signal class and combined to form a static spectrum representing the noise floor that is then subtracted from the input signal. Subtraction of identified prior sounds or background sounds is achieved by subtracting the spectral components from the input signal prior to any of the inhibitive or saturation processing steps described above, and/or by modifying the filterbank gains or the value of the saturation threshold so that signal amplitudes belonging to known components of the signal are not saturated. If the non-recognised input signal or components change over time, then the background signal class is updated accordingly. If there is substantially no saturation after adjustments of the filter gains or the saturation threshold, then the amplitudes of the non-recognised signal template(s) will therefore be reduced accordingly, leading to a decrease in the saturation threshold or an increase in the filterbank gains so that a dynamic balance is maintained over time.

- 23 -

Finally, when a template is positively identified (*i.e.*, the spectral time slices of the input signal are considered to be 'recognised', meaning that the score(s) generated for that template meets the criteria defined for positive recognition (*e.g.*, the certainty of recognition exceeds a specified threshold value)), the differences between the spectral time
5 slices of the input signal and those of the template can be determined. These differences can be used to identify secondary or subordinate signal characteristics such as speech accents, in the case of an input signal representing speech. Once identified, any such differences can be scaled by a weighting factor and added to the original templates. This enhances the recognition accuracy according to the most recent exemplars in the input
10 signal. The rate at which the templates are modified depends on the weighting factor, which is a user-defined value between zero and one.

In an alternative embodiment, the signal recognition process uses an alternative signal process 1600, as shown in Figure 16, instead of the signal process 300 of Figure 3. The
15 same initial steps 304, 306, and 308 are used to filter, rectify, and integrate the signal in the same manner as described above.

At step 1602, signal onsets are detected in the integrated filter outputs from step 308 by differentiating across each time step and then integrating across the number of filter
20 channels that approximately corresponds to a doubling of filter channel centre frequency (as described in Australian Patent Application No. 2012904074, entitled *A signal process and a signal processing system*, the entirety of which is hereby expressly incorporated herein by reference).

Some forms of input signal, including sound recordings, have a very high dynamic range that can degrade the recognition performance of the recognition process. To compress the dynamic range to a lower level, at step 1604 a noise floor is added to the integrated filter
25 outputs (*ints*) from step 308, and then the logarithm of the result is determined and the resulting values are shifted to positive values, as follows (using default values for the user-
30 configurable scale factor and offsets):

- 24 -

$$lints = \log_{10}(ints + 0.01) + 2$$

In practice, it has been found that lateral inhibition is less effective when signal amplitudes are low, and consequently is more effective if performed after saturation (rather than
5 before saturation, as described above for the first embodiment and as shown in Figure 3). To improve the effectiveness of lateral inhibition, a lateral inhibition step as described above is applied to the saturated output of step 1606 over a user-configurable number (typically 10-30) of time windows following each detected onset. The (user-configurable) inhibition weighting value is 0.3 by default in this embodiment.

10

Since the input signal amplitude is unknown, multiple saturation thresholds are used in this embodiment. An array of saturation thresholds is created that covers the expected or maximum dynamic range of the input signal (*e.g.*, an array of values from 0.01 to 1.0 in steps of 0.02, or, in the case of a sound recording, the known or estimated dynamic range
15 of the recording device can be used).

The resulting templates contain values resulting from the application of each saturation threshold across spectro-temporal domains for the signal after each onset, and are stored as a four dimensional array or matrix of j onsets, k saturation threshold levels, n channels, and
20 m time windows. A value of 0.1 is then subtracted from all these templates to generate overall inhibitory fields.

At step 1612, templates generated for each saturation threshold level are consolidated according to their similarity. To do this, each template generated using the same saturation
25 threshold level is compared to a randomly selected one of the templates (referred to hereinafter for convenience as a 'reference' template) for each level by taking the positive amplitude values in each time window (*i.e.*, same value of m) as test signals. In other words, a test signal is generated from each template and is compared to the corresponding amplitude values in the same time window of the reference template. The comparison is
30 made using the score calculation process described above, except that in this embodiment all time slices of each candidate 'test' template are compared with all equivalent time slices

- 25 -

in the reference template, regardless of whether previous time slices reached the decision threshold. Thus the positive amplitudes in each time window of the test signal are cross multiplied with the corresponding (positive and negative) amplitudes in the equivalent window of the reference template, and the results are summed (in other words, the dot product of the two vectors is determined). A match is deemed to be found if the dot product exceeds the (user-configurable) decision threshold value.

Having determined, for each saturation threshold, whether each of the templates matches the reference template, a template is selected for merging with the reference template if the percentage of time slices for which there was a match is greater than a user-configurable similarity threshold (*e.g.*, at least 70% of windows matched). Once the templates have been selected for merging, a merged template for each corresponding saturation threshold is generated as the average of all the selected templates and the reference template for the same saturation threshold. The merged templates for each saturation level are stored with the indices of the templates that were used to generate that merged template. For example, at the lowest saturation threshold level (see, for example, the graph for saturation level 1 in Figure 17) all the templates for a range of phonemes may merge, resulting in one template associated with all phoneme indices (classes, see Figure 18). This merged template can be used to recognize the presence of a voice at the pitch that the phonemes were spoken, but not what the voice said.

At higher saturation levels (such as saturation level 9 in Figure 17) relatively few templates merge, resulting in a set of distinct templates associated with only one or two phonemes each. If it is desired to recognize the presence of a voice regardless of what it is saying, then the user can choose the saturation level with the fewest templates (preferably one template). In contrast, if it is desired to recognize what is being said, the user may choose to find the saturation level with the greatest number of templates (preferably the same number of templates as phonemes used to train the system). If multiple saturation levels have the same number of templates as phonemes, then the user would choose the highest of these saturation levels, as this will contain the least common information. Alternatively, the saturation level that provides the lowest percentage of templates associated with more

- 26 -

than one phoneme can be determined. Figure 18 shows a merged template at saturation level 1 for all the phonemes used in the second Example described below.

In practice, the first step in using this embodiment of the recognition process is to
5 determine which aspect of the input signal is to be recognized, because the templates generated for different saturation levels provide the best recognition accuracy for different aspects of an input signal. Consequently, the memory templates from just one saturation level that are best associated with the particular aspect of the signal that the user wishes to discriminate are used. For example, if only the pitch of a voice is required, then templates
10 may be used from a level that discriminates pitch well but not phonemes, whereas if the user wishes to distinguish both pitch and phoneme, then templates from a different level may be used. In general, an appropriate set of templates (corresponding to one saturation level) is one whose number of member templates is closest to the expected number of classes desired to be recognised (*e.g.*, the number of phonemes at each pitch).

15

Since it is possible to determine how many channels are saturated in each memory template, a unique decision threshold value can be calculated for each time window in each template. This avoids the problem of templates with more activation being more likely to activate. In one embodiment, the decision threshold value is generated as follows:

20

$$dec(p,r) = b + m * \sum y$$

where *dec* is the decision threshold, *p* is the template index, *r* is the time window index, and *y* represents the positive amplitude values in each channel. The value of the offset *b*
25 can be selected as required to lift the decision threshold value above the noise level, and the weight *w* can be selected to increase the threshold dependency on the number of saturated channels in the template. Typical values of *b* and *w* are 0.2 and 0.4, respectively.

Once the decision threshold values have been determined, the input signal is then
30 processed as described above for the training signals to create discrete analysis frames across multiple saturation thresholds after each onset. In each time window after onset (for

- 27 -

- the length of the template), recognition is determined based on the maximum dot product of each template and the positive amplitude values of the processed input signal over all saturation levels of the input signal. The maximum dot product is then compared to the generated decision thresholds. This makes recognition independent of the input signal
- 5 amplitude and independent of any noise in the input signal at levels below the saturation threshold used by the recognition process. A template is activated if the corresponding number of matches exceeds a user-configurable activation threshold (*e.g.*, at least 50% of time windows matched)
- 10 The signal recognition processes described herein can be used for defence, border protection, security, police services and similar applications by training the process with a library of sounds of interest. The library may be small or large, depending on requirements. The process can operate on very small mobile computing platforms such as mobile phones and can use text messaging to send alerts. The process can be incorporated into a video
- 15 surveillance system (particularly with network cameras that usually include microphones). In this application, recognition of a sound of interest (such as breaking glass, for example) can be used to generate an alert to check the particular camera's view. This enables many more cameras to be actively monitored than is currently possible.
- 20 The signal recognition processes described herein can also be used for environmental noise monitoring. Noise annoyance is the greatest source of complaint in many major urban centres in the industrial world. Currently, no technology exists that can differentiate the sources of noise that contribute to an overall noise level. Different noise sources have different spectro-temporal properties and different socio-political effects that greatly
- 25 influence their acceptability for certain sectors of the community. The signal recognition processes described herein can be used to document the level and rates of occurrence of individual noise sources, and these can be used to simulate rates of noise annoyance.

EXAMPLE I

The signal recognition processes described herein provide a unique approach to music transcription and recognition in that they recognize chords that contain pitches, rather than trying to segregate and determine the pitch height of individual pitches in chords. This allows the temporal order and dynamics of sequences of chords and pitches to be transcribed. The identity of musical excerpts can then be determined by comparison to a library of digitized musical scores.

The signal recognition process of Figure 2 was applied to the recognition of musical sounds. Specifically, signal templates were generated from training signals representing single pitches and 2-pitch chords of seven equal amplitude harmonics at each interval (1-11 semitones), with the lower pitch set at each centre frequency of a Gammatone filter bank. The training and input signals were synthesized and normalized to maximum amplitudes. The recognition process was initially optimized to determine optimum values of the six process parameters using test input signals at the same frequencies as the training signals to enable unambiguous evaluation of correct recognition rates. Using the determined parameter values, the signal recognition process was then tested using single pitches and 2-pitch chords of seven equal amplitude harmonics that were synthesized at each interval (1-11 semitones) and notes of the western scale from 110 Hz (the note A2) to 440 Hz (A4). The frequencies of these input signals were randomly distributed with respect to the frequencies used for the training signals.

To demonstrate the contribution of lateral inhibition and saturation processes on the success rates of recognition, the four relevant process parameters were initially set at neurobiologically plausible values that produced reliable model outputs. A 300 channel Gammatone filter bank ($N = 300$) with center frequencies between 50 and 5000 Hz was applied to the first 50 ms portion of each input signal waveform. The outputs of the filter channels were half-wave rectified to approximate hair-cell transduction, and then integrated over 10 ms windows. A more accurate hair-cell model was unnecessary, because integration over the 10 ms window incorporates the temporal dynamics of both slow and fast refractory auditory nerves. The width of the running frequency integration window for

- 29 -

lateral inhibition was set to 15 filter channels ($I_w = 15$), and the inhibition strength factor I_s was set at 0.055, so that the minima of the inhibition side lobes about each peak were about 8% above and below the peak frequency with amplitudes of -2 to -10% of the peak amplitude. The saturation density was set at 0.24 (the proportional of channels driven to saturation, $S_d = 0.24$) which resulted in saturation thresholds (S_t) at of around 4% of the peak amplitude.

Figure 6 shows examples of spectral templates generated using the initial conditions described above and representing (i) linear excitation alone (dot-dash line 602), (ii) linear excitation with lateral inhibition (solid line 604), and (iii) saturated excitation with lateral inhibition (dotted line 606). Correct classifications were defined as trials in which the maximally activated template was the closest possible template to the target stimulus. Table I below shows that the percentage of correct classifications by the signal recognition process increased when both lateral inhibition and saturation process steps are included, thereby confirming the importance of these process steps.

15

Processes	% Correct
Linear excitation	29
Saturation only	55
Inhibition only	89
Saturation & Inhibition	100

TABLE I: Percentage of correct classifications of training stimuli with and without lateral inhibition and saturation.

For a given temporal window, the four free recognition process parameters (N , I_w , I_s and S_d) were then varied to find the extent of optimal performance. The overall strength of inhibition is affected by the combined contribution of the inhibition strength and the inhibition width, and when both values are large, all templates can be inhibited. Figure 7 (a) to (c) are three-dimensional plots showing the recognition accuracy for the training signals as a function of I_w and I_s for three respective spectral resolutions (*i.e.*, numbers of

20

- 30 -

filterbank channels, N being 22, 50, and 300 channels, respectively) at a saturation density, S_d of 0.24. As expected, the recognition rates fall to close to zero at high values of both inhibition variables. However, a diagonal plateau of very high recognition rates is observed across corresponding values of the two inhibition variables and independent of the number N of filter channels. At lower levels of inhibition, recognition rates decline to the level observed for templates with saturation only, demonstrating the enhanced selectivity due to inhibition by off-frequency excitation. These data indicate that the signal recognition process is robust across a broad range of these process parameters, and can operate effectively with numbers N of input channels similar to those used in cochlea implants.

10

To demonstrate the sensitivity of the signal recognition process to S_d and I_w , these parameters were systematically varied across a wide range of possible values to produce the four surface plots shown in Figure 8 at high and low values of N and I_s . The optimal region where recognition rates exceed 80% corresponds to S_d values of 0.15 – 0.35. Thus saturation density S_d is an insensitive parameter with a broad optimal region across varying numbers of channels and inhibition widths I_w for stimuli comprising equal amplitude harmonics. The minimum value of inhibition width I_w to produce two flanking inhibitory sidebands is three. Recognition rates exceed 80% correct for $I_w = 3$ when saturation density S_d was within its optimal range, and remained high for inhibition widths I_w up to about ten, regardless of the number of channels N . At the higher inhibition strengths I_s , recognition rates decreased more rapidly with increasing inhibition width I_w until all templates were fully inhibited in the trials with $N = 300$. These results suggest that the system performs best over the widest range of input values when inhibition width I_w is set to as small a value as practical.

25

Amplitude saturation is likely to be more critical when the amplitudes of harmonics can vary, as occurs in natural sounds. The recognition process can successfully recognize and prime the pitch of individual harmonic complexes with missing fundamentals or missing odd harmonics. TABLE II shows the recognition accuracy for chords of all intervals above the pitches of 110 Hz and 330 Hz when the odd harmonics of the component complexes were reduced in amplitude.

30

- 31 -

Frequency	Amplitude of odd harmonics					
	1	0.8	0.6	0.4	0.2	0
110 (Hz)	100	100	92	83	75	67
330 (Hz)	100	100	100	100	92	92

TABLE II: Recognition accuracy for complexes of varying odd harmonic amplitudes; with $N=300$, $S_d=0.1$, $I_w=5$, $I_s=0.095$.

- 5 For the higher frequency chords used in this example (330 Hz), recognition rates remain high even when the odd harmonics were completely removed from the complexes. The saturation density value of 0.1 used in this experiment set the saturation threshold at values between 2 and 6% of peak amplitude for chords of harmonic complexes. Odd harmonics with amplitudes greater than these threshold values were saturated. The recognition rates
- 10 dropped when odd harmonic amplitudes were reduced to lower levels, which is probably due to some harmonic amplitudes dropping below the saturation threshold after spectral inhibition. Fewer filter channels are excited by individual harmonics at higher frequencies (see Figure 6), so the saturation was generally lower for the 330 Hz chords. This caused more low amplitude harmonics to be saturated and led to higher recognition rates for the
- 15 330 Hz chords.

The optimal ranges of values for the free parameters of the recognition process were determined above using training signals that were based on the filter channel frequencies. The robustness of the recognition process was tested using a new set of 300 stimuli

20 comprising all 11 music intervals and individual pitches of the western chromatic music scale between the notes A2 (110 Hz) and A4 (440 Hz). The frequency difference between filter channels increases with channel frequency according to the equivalent rectangular bandwidth (ERB) scale, whereas the notes of the western musical scale are at logarithmic frequencies that are not aligned to the frequencies of these filter channels. In these trials,

25 correct recognition was defined as the maximal activation of the template for the correct stimulus type (single pitch or particular interval) at either of the two nearest filter channels.

- 32 -

Figure 9 is a graph of measured recognition accuracy for all the new test stimuli as a function of the number N of filter channels, and shows that the recognition rate remained above 95% when at least 100 filter channels were used. This indicates that the recognition process is not sensitive to small frequency perturbations of up to 8 Hz (*i.e.*, 1/2 the maximum frequency difference between adjacent channels when $N = 100$). This frequency difference is about 3% of the frequency of the lower pitch in the stimulus, or around one half of 1 semitone. As the number of channels decreased from about 60 to 20, the recognition rate decreased approximately linearly to 25%. While low, this recognition rate is still significantly above chance performance of 0.3%, and represents the maximum rate of recognition for musical chords that would be predicted for a cochlea implant recipient with about 20 active channels. Overall, these results demonstrate that the recognition process is robust to frequency perturbations when the number of channels is greater than 100.

15

The results described above have not included temporal inhibition. Having determined the optimal process parameters, the full signal recognition process (*i.e.*, with temporal inhibition) was tested using these optimal settings. Signal templates were generated in consecutive 10 ms integration windows for each 2-pitch chord with amplitude modulation (AM) by a full-wave rectified sine wave starting in cosine phase at 5, 10, 15 and 20 Hz.

To avoid unnecessary computations, templates were generated with the lower pitches at 172 and 307 Hz for each semitone interval 1–11, these frequencies and intervals being representative of the behaviour of the recognition process at all other frequencies. Test stimuli were synthesized for each chord with lower pitches at 175 Hz (F#3) and 311 (D#4), amplitude modulated at frequencies of 1 to 25 Hz. To evaluate the ability of the recognition process to discriminate between various AM frequencies, the level of activation of each template at both pitches was recorded at each time step for each test stimulus.

30

- 33 -

Figure 10 shows the effects of temporal inhibition and saturation on one channel of a template generated from a training signal that was amplitude modulated by a half-wave rectified sine wave starting in cosine phase. The dot-dash line represents the amplitude of the frequency channel as a function of time as output by the Gammatone filterbank. After saturation (solid line), the template becomes very sensitive to onsets and offsets of the amplitude modulation as it crosses the saturation threshold, leading to sharp edges of the excitatory regions. Regions of temporal inhibition were then generated immediately after each excitatory region by subtraction of the weighted sum of activation in the preceding time steps, as described above (with the resulting data indicated by the dotted line). The onset spectral slice of each region of excitation does not receive temporal inhibition, and consequently has larger amplitudes. This occurs for all templates, and so will not affect the reliability of recognition based on initial spectral information.

Figures 11 to 14 are graphs showing the degree of activation over time of each of four templates representing a 172 Hz harmonic complex with 0, 5, 10 and 20 Hz amplitude modulation (AM), respectively, and where the input signal (or 'stimulus') to be recognised was a harmonic complex, also at the same frequency of 172 Hz and with 0, 5, 10 and 20 Hz AM for the respective four Figures. The optimal temporal inhibition integration time and strength for these signals were found to be respectively 3 ms and 0.125 by systematic variation over the possible range values. In each case, the activation of the correct AM template can be seen to increase over time relative to templates at other AM frequencies. The mechanism for the template discrimination can be understood by examining the results for the non-modulated stimulus in Figure 11. Initially all templates are equally activated, because they share identical spectral information at onset. The activation of the 20 Hz AM template drops first due to the presence of input signal excitation in a region of temporal inhibition in the template after the first positive phase of AM (around 25 ms). Other templates with longer AM periods decrease in activation over the next 25 ms. For AM stimuli (Figures 12 to 14), the activation of the non-modulated template stopped increasing after the loss of stimulus excitation until the next positive phase of the stimulus. The AM templates increased in activation more than the non-modulated template due to higher template levels in the early stages of each positive phase after the onset (Figure 10).

- 34 -

For a certainty of recognition (being the ratio of activation between the strongest and second strongest activated templates) of 1.3, the recognition process required around 220 ms of the input signal to distinguish between the non-modulated tone and the 10 Hz AM tone, and around 50 ms for the 20 Hz AM tone. However the 5 Hz AM tone would not be distinguished from the non-modulated tone within 300 ms. In other words, the 5 Hz AM stimulus would be perceived as the original tone beating in amplitude. The frequency range of AM that corresponds to the perception of roughness rather than beating begins at around 20 Hz. This is the AM frequency at which the recognition mechanism is able to robustly distinguish AM tones within 50 ms, the time in which recognition must occur to follow speech. These findings suggest that roughness becomes a perceptual attribute associated with a specific template (or auditory identity) due to the temporal dynamics of recognition mechanisms when AM frequencies are greater than about 20 Hz. These findings accord with the common experience that beating sounds are not distinguished as different sound sources by their rate of beating, whereas roughness may be used to distinguish between sources.

Figure 15 shows the activation of the AM templates used in Figures 11 to 14 by harmonic complex stimuli at varying AM rates 100 ms after onset. All the templates for both 175 Hz and 311 Hz stimuli were maximally activated by stimuli at their respective pitch and AM frequency within 100 ms of stimulus onset. Furthermore, Figure 15 shows that templates with AM frequencies equal to or greater than 15 Hz were well distinguished from the non-modulated complex across a range of AM frequencies, indicating that AM frequencies above 15 Hz can be recognized by a sparse set of AM templates. In other words, just a few sound identities would be associated with a wide range of AM frequencies.

The results described above demonstrate that the signal recognition process was able to correctly recognize all 2-pitch chords and single pitches of 7-harmonic complexes across the range of pitches from 110 Hz (the note A2) to 440 Hz (A4). It was also able to robustly recognize signals with reduced amplitudes of odd harmonics and AM frequencies greater than 15 Hz within 100 ms. This is the minimum performance required to match the

- 35 -

performance of well-trained musicians. Overall recognition rates based on template matching of spectral information increased from 29% in the absence of both inhibition or saturation processes to 100% when these processes were included, highlighting the importance of these processes.

5

In the frequency domain, the signal recognition process has only four free input variables, and is robust to variations in these variables across a wide range. In particular, recognition rates are not diminished by the reduction of the number of input frequency channels from 300 to around 20. Significantly, around 20 channels are sufficient to successfully encode speech

10 signals in cochlea implants.

A further two parameters (temporal inhibition integration time and strength) are used to define temporal inhibition. Since the temporal integration window of 10 ms is quite long compared to the rates of AM, there was only a small range of possible values for these
15 variables. The signal recognition process was found to be robust to small frequency differences between long-term memory templates and stimuli and between differences in the rates of AM.

- 36 -

EXAMPLE II

The recognition process using the alternative signal process of Figure 16 was tested using 32 phonemes recorded in a single data file at the same pitch by the same speaker using a laptop computer microphone. The phonemes are a representative sample of all Australian English phonemes. Every vowel sound was included with the consonant “h”, and every consonant was included with the vowel “a” as in “cat” (see Table 3 below). Table 4 provides the values of all of the parameters used in this example.

Table 3. Phonemes used to test the recognition process

Vowels	Consonants
Hi (hid)	Ma (mat)
He (hair)	Na
Ha (hard)	Pa
Ha (had)	Ta
Ho (hot)	Ca
Her	Ba
Who	Da
High	Ga
Ho	Cha
Hay	Ja
How	Fa
Hoy	Tha
	Sa
	Sha
	Va
	Za
	Wa
	Ya
	Ra
	La

Table 4. Initial parameters used in Example II, except where explicitly varied in the examples below.

Parameter	Value	Parameter	Value
Sampling Frequency	11,025 sec ⁻¹	Resolution of minimum saturation thresholds	0.02
Number of Channels	50	Lateral and temporal inhibition strength	0.3
Lowest Frequency Channel	70 Hz	Lateral inhibition width	1 channel
Highest Frequency Channel	5,500 Hz	Temporal inhibition window	6 msec
Temporal integration window	10 msec	Template length	200 msec
Onset threshold	1.3	Activation threshold for template merging	0.5
Onset integration width	10 channels	Decision threshold for template merging	0.7
Lowest minimum saturation threshold	0.01	Bias for computing recognition threshold [b]	0.2
Highest minimum saturation threshold	3.0	Template amplitude weight for computing recognition threshold [w]	0.4

- 5 Initially, the minimum number of channels required to achieve maximal recognition of phoneme classes was evaluated. Recognition rates increased from 20% with 20 channels, to 94% with 35 channels, and finally 100% for 50 - 300 channels. Since 50 channels were adequate for 100% recognition, this value was used for the remainder of the validation.
- 10 To test the loudness independence of the recognition process, the phoneme recording was varied in amplitude. Figure 19 shows that the recognition rates remained above 90% over a range of 80dB, which is greater than the range of loudness usually encountered in modern

- 38 -

city environments. This amplitude range was achieved with 50 levels of saturation thresholds between 0.01 and 1. The number of saturation threshold levels can be reduced if a smaller amplitude range is required.

- 5 Pink noise was added to the recording to test the robustness of the recognition process to noise when using templates created without noise. Figure 20 shows that recognition rates remained above 90% for signal to noise ratios (SNR) as low as 2.5 dB, given reliable onset detection. However, onset detection failed on 33% of actual onsets at SNR of 2.5 dB, resulting in lower overall recognition rates. Onset detection rates of 100% have been
- 10 obtained for speech in pink noise at a SNR of -4 dB using this process with an integration time window of 1 msec rather than the 10 msec windows used in this example. Consequently, it can be expected that 90% overall recognition rates could be achieved at a SNR of 2.5 dB. This is considered to be substantially better than existing methods that begin to fail at a SNR of 10 dB. This high rate of recognition in the presence of noise is
- 15 due to use of templates that capture amplitude information at minimum amplitude thresholds. So whenever noise levels are below the minimum amplitude threshold, they have no effect on the recognition mechanism.

- Finally, to test whether phoneme recognition is independent of pitch, a new test file was
- 20 created using the phonemes "ma", "ba" and "sa", recorded at each note of a major scale over an octave commencing at 96 Hz. Eighty-eight percent of the phonemes were correctly recognized at the correct pitches.

- Many modifications will be apparent to those skilled in the art without departing from the
- 25 scope of the present invention.

- 39 -

APPENDIX – Sample Matlab Code Listing

```

TRAIN

5  clear all

    numChannels=50;
    fs=11025;

10  sig=wavread('train.wav');

    %frame signal
    inttime=0.01;
    intsamp=floor(inttime*fs);
    %time length of integration window
15  lensig=length(sig);
    noint=floor(lensig/intsamp);

    %transform signal into features

20  ints(numChannels,noint)=zeros;

    lowFreq=70;
    highFreq=5500;

25  fcoefs=MakeERBFilters(fs,numChannels,lowFreq);
    data = ERBFilterBank(sig, fcoefs);
    sampleRate=fs;

30  clear sig

    EarQ = 9.26449;
    minBW = 24.7;
    order = 1;
    % Glasberg and Moore Parameters

35  cfArray = -(EarQ*minBW) + exp((1:numChannels)'*(-log(highFreq +
    EarQ*minBW) + ...
        log(lowFreq + EarQ*minBW))/numChannels) * (highFreq +
    EarQ*minBW);

40

    for n=1:numChannels
        for i=1:lensig;
            if data(n,i)<=0
45                data(n,i)=0;
            end
        end
    end

50

```

- 40 -

```

    for n=1:numChannels
        for i=1:(noint-1);
            ints(n,i)=sum(data(n,((i-1)*intsamp+1):(i*intsamp)));
        end
5    end

    clear data

    s=size(ints,2);

10    %find onsets

    onthres=1.3;
    j=1;
15    onset=[];
    sints=[];
    dsints=[];

    for n=1:numChannels-11
20        sints(n,:)=sum(ints(n:n+10,:),1);
    end

    for i=2:s
        dsints(:,i)=sints(:,i)-sints(:,i-1);
25    end

    for i=2:s
        if max(dsints(:,i))>onthres;
30            onset(j)=i;
            j=j+1;
            dsints(:,i:(i+50))=zeros;
        end
    end

35    clear sints dsints

    nonset=length(onset);

    %flanking inhibition then saturation
40    ints=ints+0.01;
    ints=log10(ints)+2;
    thresh=0.01:0.02:1.3;
    numthresh=length(thresh);
45    ICsat(1:numthresh,1:numChannels,1:noint)=zeros;
    inhints(1:numChannels,1:noint)=zeros;

    inhstr=0.3;

```

- 41 -

```

clear ints

5  for k=1:numthresh
    for n=1:numChannels
        for i=1:noint
            if ints(n,i)-thresh(k)>0;
                ICsat(k,n,i)=1;
10         end
        end
    end
end

15  for k=1:numthresh
    for n=2:numChannels-1
        for i=2:noint
            ICsat(k,n,i)=ICsat(k,n,i)-ICsat(k,n-1,i).*inhstr-
20         ICsat(k,n+1,i).*inhstr-ICsat(k,n,i-1).*inhstr;
        end
    end
end

clear inhints ints

25  %create analysis windows
    lentem=20;
    tem(1:nonset,1:numthresh,1:numChannels,1:lentem)=zeros;
    ICtem(1:nonset,1:numthresh,1:numChannels,1:lentem)=zeros;
30  mtem(1:nonset,1:numthresh,1:numChannels,1:lentem)=zeros;

    for j=1:nonset
        tem(j,:,:,:)=ICsat(:,:,:onset(j):onset(j)+lentem-1);
    end

35  tem=tem-0.1;

    %find self similar
    dec=0.5; %decision threshold
40  acttem(1:nonset,1:nonset,1:numthresh,1:lentem)=zeros;
    sumact(1:nonset,1:nonset,1:numthresh,1:lentem)=zeros;
    decs(1:nonset,1:nonset,1:numthresh,1:lentem)=zeros;

```

- 42 -

```

for j=1:nonset
    for i=1:lentem
        for k=1:numthresh
            x=squeeze(tem(j,k,:,i));
5            for n=1:numChannels
                if x(n)<0
                    x(n)=0;
                end
            end
10            for m=1:nonset
                y=squeeze(squeeze(tem(m,k,:,i)));
                decs(j,m,k,i)=sum(x.*y);
                if decs(j,m,k,i)>=dec
                    acttem(j,m,k,i)=1;
15                end
                sumact(j,m,k,i)=sum(acttem(j,m,k,1:i));
            end
        end
    end
20 end

sumall=squeeze(sumact(:,:,:,lentem));
sumall=sumall/20;

25 %merge templates
tempind(1:numthresh,1:nonset,1:nonset)=zeros;
ntempind(1:numthresh,1:nonset,1:nonset)=zeros;
templates.ind='indices';
templates.temps='templates';
30 templates.num='num';

for k=1:numthresh
    inds(1:nonset,1:nonset)=zeros;
    count(k)=1;
35    for j=1:nonset
        m=max(sumall(j,:,k));
        a=find(sumall(j,:,k)>0.7*m); %must be more than 70% activated to
        be included in template
        inds(j,a)=ones;
40    end
    tempind(k,1,:)=inds(1,:);
    for p=1:count(k)
        for j=1:nonset
            if inds(j,j)>0;
45            if tempind(k,p,j)>0;

tempind(k,p,:)=squeeze(tempind(k,p,:))+squeeze(inds(j,:));
            else
                count(k)=count(k)+1;
50                tempind(k,count(k),:)=squeeze(inds(j,:));
            end
        end
    end
    end
55    for p=1:count(k)
        if sum(tempind(k,p,:))>0

```

- 43 -

```

ntempind(k,p,:)=tempind(k,p,:)./sum(tempind(k,p,:));
template(1:numChannels,1:lentem)=zeros;
for j=1:nonset
    temp(:,:,)=squeeze(tem(j,k,:,:));
5   template(:,:,)=template(:,:,)+ntempind(k,p,j).*temp(:,:,);
    end
    rr=find(tempind(k,p,:)>0);
    ss=squeeze(sumall(p,rr,k));
10   templates(k,p).ind=[rr,ss'];
    templates(k,p).temps=template;
    templates(k,p).num=count(k);
    end
end
15 end

for k=1:size(templates,1)
    if templates(k,1).num > 0
        notemp(k)=templates(k,1).num;
20   end
end

bb=find(notemp==max(notemp));
25 ind1=max(bb);

fintemp(1).num=templates(ind1,1).num;

for p=1:templates(ind1,1).num
30   fintemp(p).ind=templates(ind1,p).ind;
    fintemp(p).temps=templates(ind1,p).temps;
end

35 save('merged_temps_pitch','fintemp');
```

- 44 -

```
TEST

clear all

5  numChannels=50;
   fs=11025;

   sig=wavread('test.wav');

10  inttime=0.01;           %time length of integration window
   intsamp=floor(inttime*fs);
   lensig=length(sig);
   noint=floor(lensig/intsamp);

15  clear sig noise

   %transform signal into features

20  ints(numChannels,noint)=zeros;

   lowFreq=70;
   highFreq=5500;

25  fcoefs=MakeERBFilters(fs,numChannels,lowFreq);
   data = ERBFilterBank(in, fcoefs);
   sampleRate=fs;

30  clear sig

   EarQ = 9.26449;           % Glasberg and Moore Parameters
   minBW = 24.7;
   order = 1;

35
```

- 45 -

```

cfArray = -(EarQ*minBW) + exp((1:numChannels)'*(-log(highFreq +
EarQ*minBW) + ...
    log(lowFreq + EarQ*minBW))/numChannels) * (highFreq +
EarQ*minBW);
5

for n=1:numChannels
    for i=1:lensig;
        if data(n,i)<=0
10            data(n,i)=0;
        end
    end
end

15
for n=1:numChannels
    for i=1:(noint-1);
        ints(n,i)=sum(data(n,((i-1)*intsamp+1):(i*intsamp)));
    end
20 end

clear data in

s=size(ints,2);
25
%find onsets

onthres=1.3;
j=1;
30 onset=[];
sints=[];
dsints=[];

for n=1:numChannels-11
35 sints(n,:)=sum(ints(n:n+10,:),1);
end

for i=2:s
    dsints(:,i)=sints(:,i)-sints(:,i-1);
40 end

for i=2:s
    if max(dsints(:,i))>onthres;
        onset(j)=i;
45 j=j+1;
        dsints(:,i:(i+50))=zeros;
    end
end

50 clear sints dsints

nonset=length(onset);

```

- 46 -

```

%flanking inhibition then saturation

ints=ints+0.01;
ints=log10(ints)+2;
5 thresh=0.01:0.02:1;
numthresh=length(thresh);
ICsat(1:numthresh,1:numChannels,1:noint)=zeros;
inhints(1:numChannels,1:noint)=zeros;

10 inhstr=0.3;

for k=1:numthresh
15 for n=1:numChannels
    for i=1:noint
        if ints(n,i)-thresh(k)>0;
            ICsat(k,n,i)=1;
        end
20    end
end

for k=1:numthresh
25 for n=2:numChannels-1
    for i=2:noint
        ICsat(k,n,i)=ICsat(k,n,i)-ICsat(k,n-1,i).*inhstr-
        ICsat(k,n+1,i).*inhstr-ICsat(k,n,i-1).*inhstr;
    end
30 end
end

clear inhints ints

35 open merged_temps_pitch.mat;

templates=ans.fintemp;
notemp=templates(1,1).num;
lentem=size(templates(1,1).temps,2);
40 levtem=size(templates,1);
decmatlen=nonset*lentem+1;
acttem(1:notemp,1:decmatlen)=zeros;
sumact(1:notemp,1:nonset)=zeros;
dec(1:notemp,1:lentem)=zeros;
45 decs(1:numthresh,1:notemp,1:decmatlen)=zeros;
decs1(1:notemp,1:decmatlen)=zeros;
indices(1:notemp,1:nonset)=zeros;
step(1:nonset)=zeros;
cor(1:nonset)=zeros;

50 for p=1:notemp
    for r=1:lentem
        tem=templates(1,p).temps;
        y=squeeze(tem(:,r));
55 for n=1:numChannels

```

- 47 -

```

        if y(n)<0
            y(n)=0;
        end
    end
    dec(p,r)=0.2+0.4*sum(y);
5    end
end

for j=1:nonset
10    i=onset(j);
    step(j)=(j-1)*lentem;
    for p=1:notemp
        tem=templates(1,p).temps;
        for r=1:lentem
15            y=squeeze(tem(:,r));
            for q=1:numthresh
                x=squeeze(ICsat(q,:,i+r-1));
                for n=1:numChannels
20                    if x(n)<0
                        x(n)=0;
                    end
                end
                if sum(x)>0
25                    decs(q,p,step(j)+r)=sum(x.*y');
                end
            end
            decs1(p,step(j)+r)=max(decs(:,p,step(j)+r));
            if decs1(p,step(j)+r)>=dec(p,r)
                acttem(p,step(j)+r)=1;
30            end
            if r==lentem
                sumact(p,j)=sum(acttem(p,step(j)+1:(step(j)+lentem)));
            end
        end
    end
35    end
end

for j=1:nonset
    [m,mm]=sort(sumact(:,j));
40    if mm(nonset)==j
        cor(j)=1;
    end
end

45    pcor=sum(cor).*100/notemp

figure(1), contourf(sumact(:,:))

```

CLAIMS:

1. A signal recognition process, including:

receiving signal data representing a signal;

filtering the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data;

setting signal amplitudes exceeding a saturation threshold to a saturation value representing reinforcement; and

applying lateral inhibition across each of the one or more other dimensions to generate, for each said other dimension, inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension.

2. The signal recognition process of claim 1, including:

applying temporal inhibition to the signal amplitudes to produce inhibitive signal amplitude values immediately following offsets of the saturated signal amplitudes.

3. A signal recognition process, including:

receiving signal data representing a signal;

filtering the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data;

applying lateral inhibition across each of the one or more other dimensions to generate inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension; and

applying temporal inhibition to the signal amplitudes to produce inhibitive signal amplitude values immediately following offsets of the saturated signal amplitudes.

4. A signal recognition process, including:
 - receiving training signal data representing one or more training signals and processing the training signal data to generate signal recognition templates using the process of any one of claims 1 to 3;
 - receiving input signal data representing an input signal to be recognised and processing the input signal data to generate processed input signal data using the process of any one of claims 1 to 3;
 - for each of the signal recognition templates, generating a corresponding recognition score quantifying correspondence between the processed input signal data and the signal recognition template.
5. The process of claim 4, including selecting, on the basis of the generated scores, at least one of the signal recognition templates as corresponding to the input signal.
6. The signal recognition process of any one of claims 1, 2, 4 or 5, including determining the saturation threshold such that a specified proportion of the signal amplitudes exceed the saturation threshold.
7. The signal recognition process of any one of claims 1, 2, 4, or 5, wherein the step of setting signal amplitudes includes, for each of a plurality of saturation thresholds, generating a corresponding set of recognition templates in which signal amplitudes exceeding the corresponding saturation threshold are set to a saturation value representing reinforcement.
8. The signal recognition process of any one of claims 4 to 7, including:
 - generating, for each of a plurality of time windows of each of said templates, a corresponding decision value based on the corresponding amplitude values of the template;
 - for each of a plurality of time windows following a detected signal onset, generating dot products of the corresponding positive amplitude values of the

- 50 -

- processed input signal data and the corresponding amplitude values of respective ones of the signal recognition templates; and
for each of the plurality of time windows following the detected signal onset, comparing a maximum one of the generated dot products with the corresponding
5 decision value for the time window, and determining whether the corresponding signal recognition template is a match to the input signal for the time window based on said comparing; and
selecting at least one of said signal recognition templates as being recognised based on the number of matches of the at least one signal recognition template to the
10 input signal.
9. The signal recognition process of any one of claims 4 to 8, including reducing the number of said signal recognition templates by combining similar ones of said signal recognition templates identified by generating scores quantifying
15 correspondence between at least some of said signal recognition templates.
10. The process of any one of claims 4 to 9, wherein a plurality of said signal recognition templates are generated for successive temporal portions of each training signal, and each recognition score is generated from a template for a corresponding temporal portion of a training signal and processed input signal data for a corresponding temporal portion of the input signal.
11. The process of any one of claims 4 to 10, wherein the received input signal represents a combination of a first signal corresponding to one of the training signals and at least one second signal overlapping with the first signal, the selected signal recognition template corresponds to a first temporal portion of the first signal, and the process further includes determining predicted first signal data on the basis of a further signal recognition template corresponding to a second temporal portion of the first signal subsequent to the first portion of the first signal,

and using the predicted first signal data to improve recognition of the at least one second signal.

12. The process of any one of claims 1 to 7, including generating one or more background templates from unrecognised temporal portions of the input signal, and using the generated background templates to improve the recognition of input signal components in subsequent temporal portions of the input signal.

13. A signal process, including:

(i) for each of a plurality of training signals, generating a set of signal templates corresponding to successive temporal portions of the training signal;

(ii) processing successive temporal portions of an input signal to generate respective processed input signal portion data;

(iii) selecting a subset of the signal templates corresponding to selected temporal portions of each training signal and processed input signal portion data representing a corresponding selected temporal portion of the input signal;

(iv) for each said training signal, processing the corresponding signal template and the selected processed input signal portion data to generate a corresponding score representing correspondence between the selected temporal portion of the training signal and the selected temporal portion of the input signal;

(v) selecting a further subset of the signal templates representing a subsequent temporal portion of each training signal and processed input signal portion data representing a corresponding further temporal portion of the input signal; and

(vi) repeating step (iv) to generate further scores for the further temporal portions.

14. The process of claim 13, wherein in step (v) only signal templates from sets of templates whose scores generated at step (iv) exceeded a threshold value are selected.

- 52 -

15. The process of claim 13 or 14, wherein step (vi) includes repeating step (iv) until the generated scores satisfy one or more predetermined criteria.
16. The process of any one of claims 1 to 15, wherein the process is substantially a real-time process.
17. The process of any one of claims 1 to 16, wherein the one or more other dimensions include frequency.
18. The process of any one of claims 1 to 17, wherein the one or more other dimensions include one or more spatial dimensions.
19. The process of any one of claims 1 to 18, wherein the signal includes an audio signal.
20. The process of any one of claims 1 to 19, wherein the signal includes a video signal.
21. The process of any one of claims 1 to 20, wherein the process is configured to recognise sounds.
22. The process of any one of claims 1 to 21, wherein the process is configured to recognise human speech.
23. A computer-readable storage medium having stored thereon processor-executable instructions that, when executed by a processor, cause the processor to execute the process of any one of claims 1 to 22.

- 53 -

24. A signal recognition system configured to execute the process of any one of claims 1 to 22.

25. A signal recognition system, including:

a signal processing component configured to:

(i) receive signal data representing a signal;

(ii) filter the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data;

5 (iii) set signal amplitudes exceeding a saturation threshold to a saturation value representing reinforcement; and

(iv) apply lateral inhibition across each of the one or more other dimensions to generate, for each said other dimension, inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension;

a training component configured to receive training signal data representing one or more training signals and to cause the signal processing component to process the training signal data to generate signal recognition templates; and

a signal recognition component configured to:

(a) receive input signal data representing an input signal to be recognised and to cause the signal processing component to process the input signal data to generate processed input signal data; and

(b) for each of the signal recognition templates, to generate a corresponding recognition score quantifying correspondence between the processed input signal data and the signal recognition template.

AMENDED CLAIMS

received by the International Bureau on 20 February 2013 (20.02.2013)

CLAIMS:

1. A signal recognition process, including:
 - receiving signal data representing a signal;
 - filtering the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data;
 - 5 setting signal amplitudes exceeding a saturation threshold to a saturation value representing reinforcement; and
 - applying lateral inhibition across each of the one or more other dimensions to generate, for each said other dimension, inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension.
 - 10
2. The signal recognition process of claim 1, including:
 - applying temporal inhibition to the signal amplitudes to produce inhibitive signal amplitude values immediately following offsets of the saturated signal amplitudes.
 - 15
3. A signal recognition process, including:
 - receiving signal data representing a signal;
 - filtering the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data;
 - 20 applying lateral inhibition across each of the one or more other dimensions to generate inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension; and
 - applying temporal inhibition to the signal amplitudes to produce inhibitive signal amplitude values immediately following offsets of the signal amplitudes.
 - 25

4. A signal recognition process, including:

receiving training signal data representing one or more training signals and processing the training signal data to generate signal recognition templates using the process of any one of claims 1 to 3;

receiving input signal data representing an input signal to be recognised and processing the input signal data to generate processed input signal data using the process of any one of claims 1 to 3;

for each of the signal recognition templates, generating a corresponding recognition score quantifying correspondence between the processed input signal data and the signal recognition template.

5. The process of claim 4, including selecting, on the basis of the generated scores, at least one of the signal recognition templates as corresponding to the input signal.

6. The signal recognition process of any one of claims 1, 2, 4 or 5, including determining the saturation threshold such that a specified proportion of the signal amplitudes exceed the saturation threshold.

5 7. The signal recognition process of any one of claims 1, 2, 4, or 5, wherein the step of setting signal amplitudes includes, for each of a plurality of saturation thresholds, generating a corresponding set of recognition templates in which signal amplitudes exceeding the corresponding saturation threshold are set to a saturation value representing reinforcement.

10 8. The signal recognition process of any one of claims 4 to 7, including:
generating, for each of a plurality of time windows of each of said templates, a corresponding decision value based on the corresponding amplitude values of the template;
for each of a plurality of time windows following a detected signal onset,
15 generating dot products of the corresponding positive amplitude values of the

- processed input signal data and the corresponding amplitude values of respective ones of the signal recognition templates; and
for each of the plurality of time windows following the detected signal onset, comparing a maximum one of the generated dot products with the corresponding
5 decision value for the time window, and determining whether the corresponding signal recognition template is a match to the input signal for the time window based on said comparing; and
selecting at least one of said signal recognition templates as being recognised based on the number of matches of the at least one signal recognition template to the
10 input signal.
9. The signal recognition process of any one of claims 4 to 8, including reducing the number of said signal recognition templates by combining similar ones of said signal recognition templates identified by generating scores quantifying
15 correspondence between at least some of said signal recognition templates.
10. The process of any one of claims 4 to 9, wherein a plurality of said signal recognition templates are generated for successive temporal portions of each training signal, and each recognition score is generated from a template for a corresponding temporal portion of a training signal and processed input signal data for a corresponding temporal portion of the input signal.
11. The process of any one of claims 4 to 10, wherein the received input signal represents a combination of a first signal corresponding to one of the training signals and at least one second signal overlapping with the first signal, the selected signal recognition template corresponds to a first temporal portion of the first signal, and the process further includes determining predicted first signal data on the basis of a further signal recognition template corresponding to a second temporal portion of the first signal subsequent to the first portion of the first signal,

and using the predicted first signal data to improve recognition of the at least one second signal.

12. The process of any one of claims 1 to 7, including generating one or more background templates from unrecognised temporal portions of the input signal, and using the generated background templates to improve the recognition of input signal components in subsequent temporal portions of the input signal.

13. A signal process, including:

(i) for each of a plurality of training signals, generating a set of signal templates corresponding to successive temporal portions of the training signal;

(ii) processing successive temporal portions of an input signal to generate respective processed input signal portion data;

(iii) selecting a subset of the signal templates corresponding to selected temporal portions of each training signal and processed input signal portion data representing a corresponding selected temporal portion of the input signal;

(iv) for each said training signal, processing the corresponding signal template and the selected processed input signal portion data to generate a corresponding score representing correspondence between the selected temporal portion of the training signal and the selected temporal portion of the input signal;

(v) selecting a further subset of the signal templates representing a subsequent temporal portion of each training signal and processed input signal portion data representing a corresponding further temporal portion of the input signal; and

(vi) repeating step (iv) to generate further scores for the further temporal portions.

14. The process of claim 13, wherein in step (v) only signal templates from sets of templates whose scores generated at step (iv) exceeded a threshold value are selected.

15. The process of claim 13 or 14, wherein step (vi) includes repeating step (iv) until the generated scores satisfy one or more predetermined criteria.
16. The process of any one of claims 1 to 15, wherein the process is substantially a real-time process.
17. The process of any one of claims 1 to 16, wherein the one or more other dimensions include frequency.
18. The process of any one of claims 1 to 17, wherein the one or more other dimensions include one or more spatial dimensions.
19. The process of any one of claims 1 to 18, wherein the signal includes an audio signal.
20. The process of any one of claims 1 to 19, wherein the signal includes a video signal.
21. The process of any one of claims 1 to 20, wherein the process is configured to recognise sounds.
22. The process of any one of claims 1 to 21, wherein the process is configured to recognise human speech.
23. A computer-readable storage medium having stored thereon processor-executable instructions that, when executed by a processor, cause the processor to execute the process of any one of claims 1 to 22.

24. A signal recognition system configured to execute the process of any one of claims 1 to 22.

25. A signal recognition system, including:

a signal processing component configured to:

(i) receive signal data representing a signal;

(ii) filter the signal data to generate filtered data representing signal amplitudes as a function of time and one or more other dimensions represented by the signal data;

5 (iii) set signal amplitudes exceeding a saturation threshold to a saturation value representing reinforcement; and

(iv) apply lateral inhibition across each of the one or more other dimensions to generate, for each said other dimension, inhibitive signal amplitude values at values of said dimension flanking dominant ones of the signal amplitudes along said dimension;

a training component configured to receive training signal data representing one or more training signals and to cause the signal processing component to process the training signal data to generate signal recognition templates; and

a signal recognition component configured to:

(a) receive input signal data representing an input signal to be recognised and to cause the signal processing component to process the input signal data to generate processed input signal data; and

(b) for each of the signal recognition templates, to generate a corresponding recognition score quantifying correspondence between the processed input signal data and the signal recognition template.

1/21

100

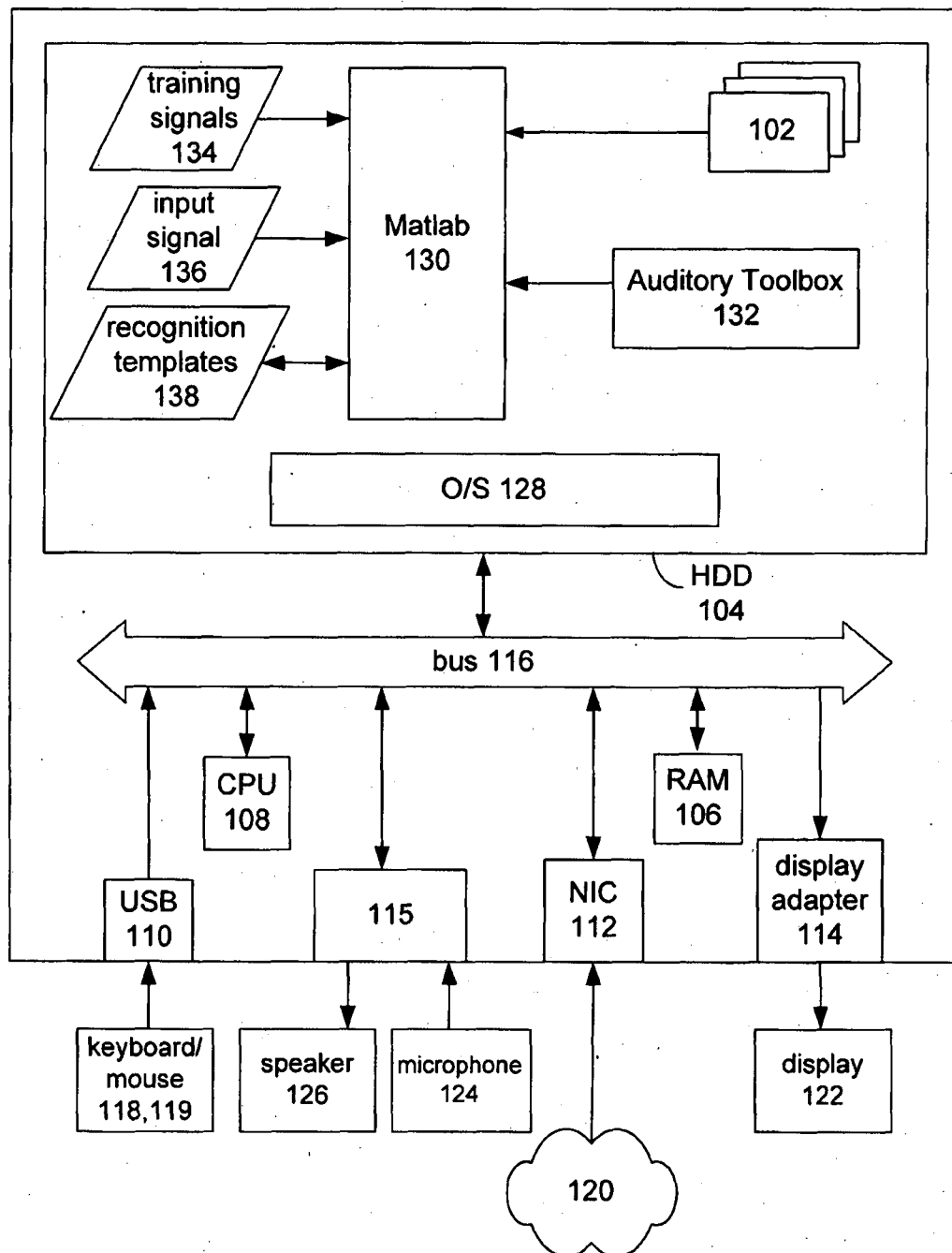


Figure 1

2/21

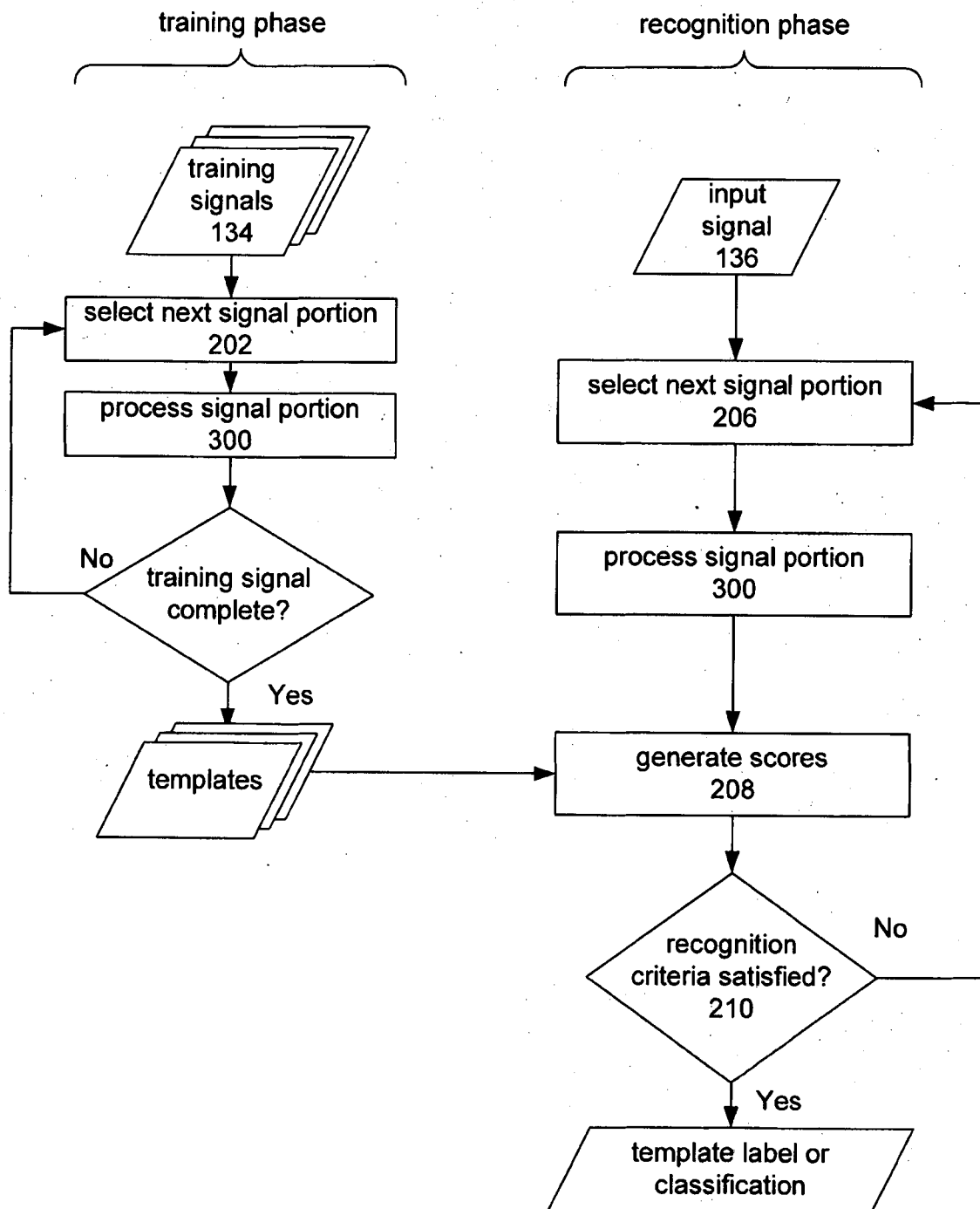


Figure 2

3/21

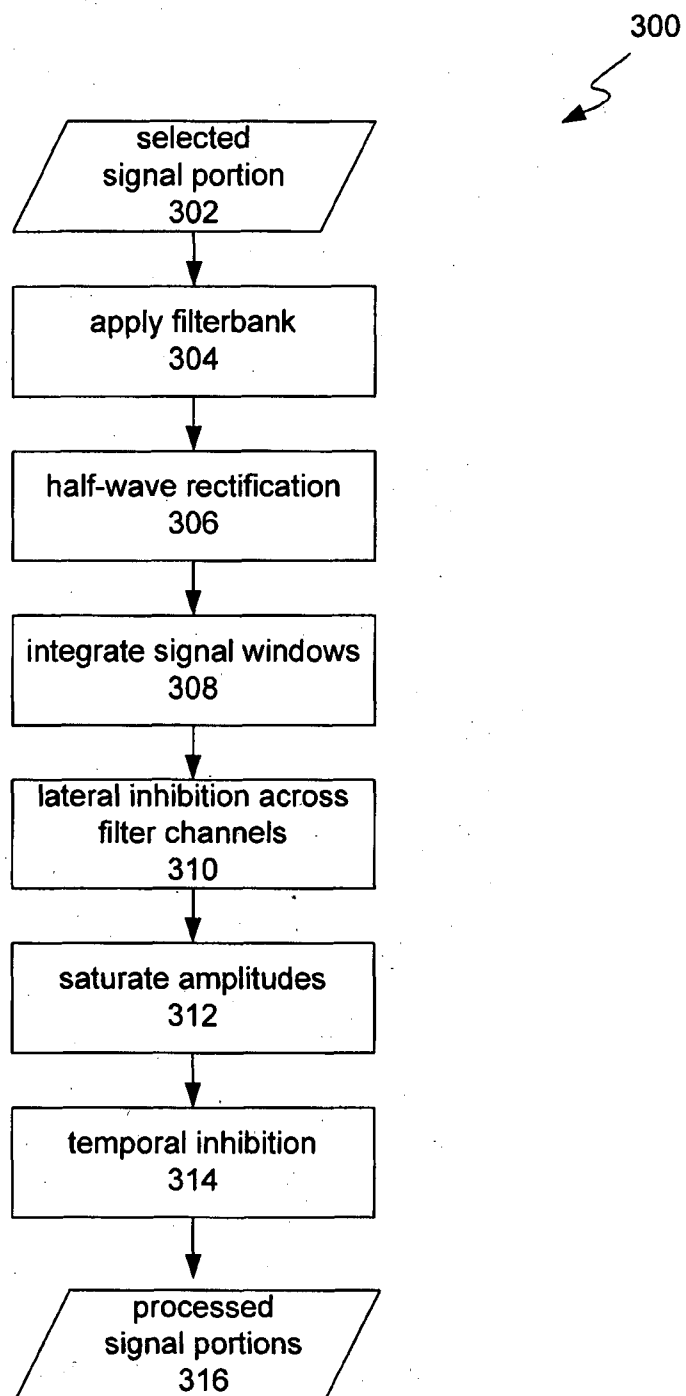


Figure 3

4/21

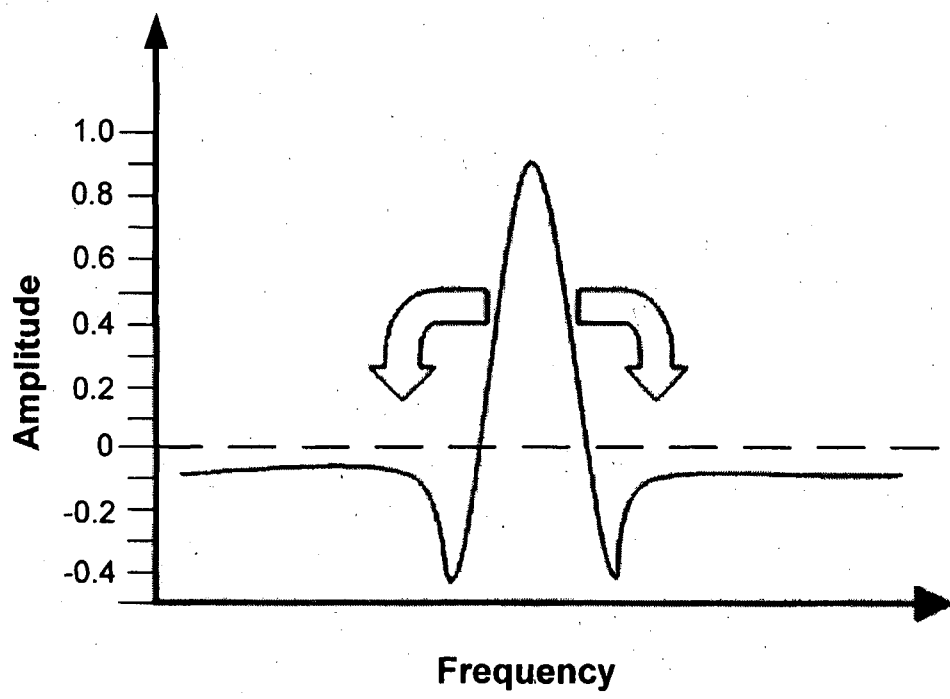


Figure 4

5/21

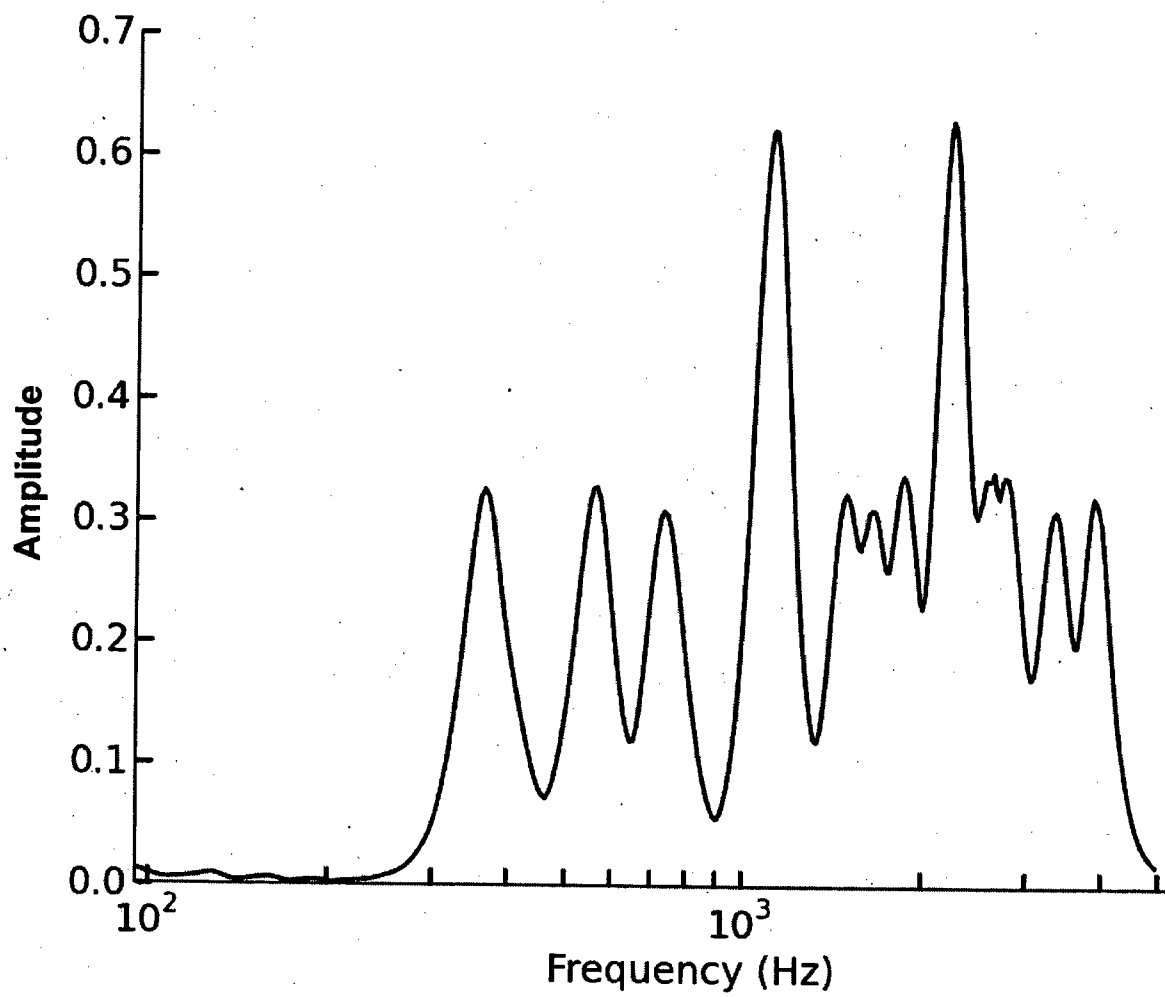


Figure 5

6/21

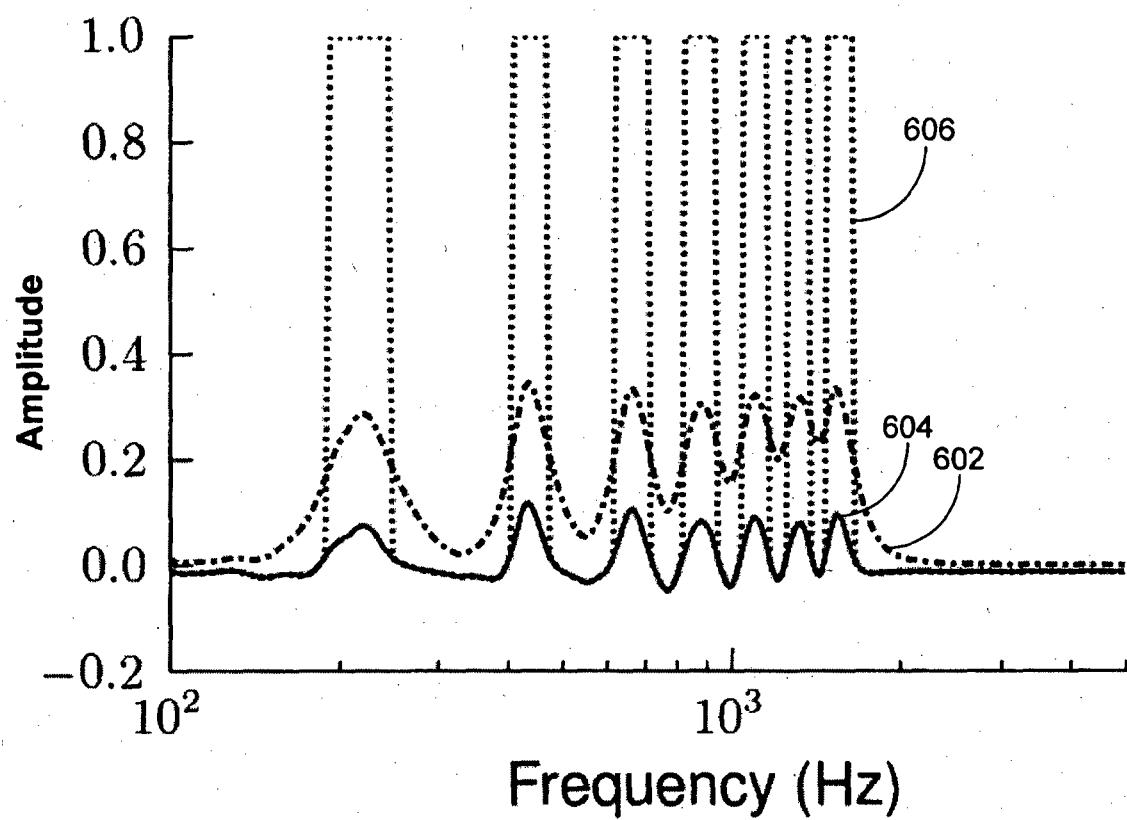


Figure 6

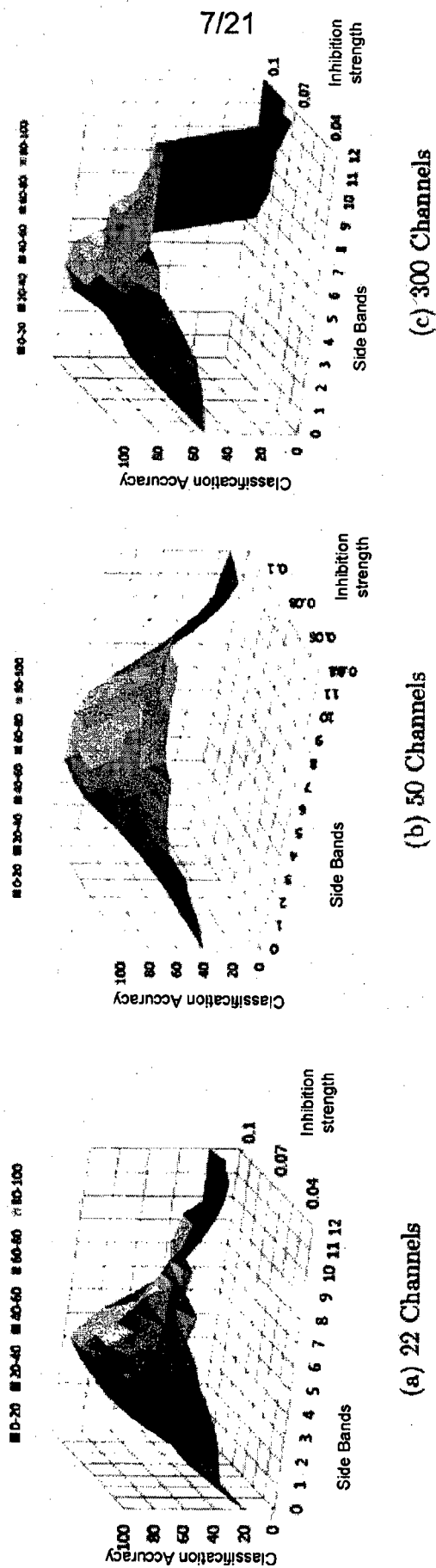
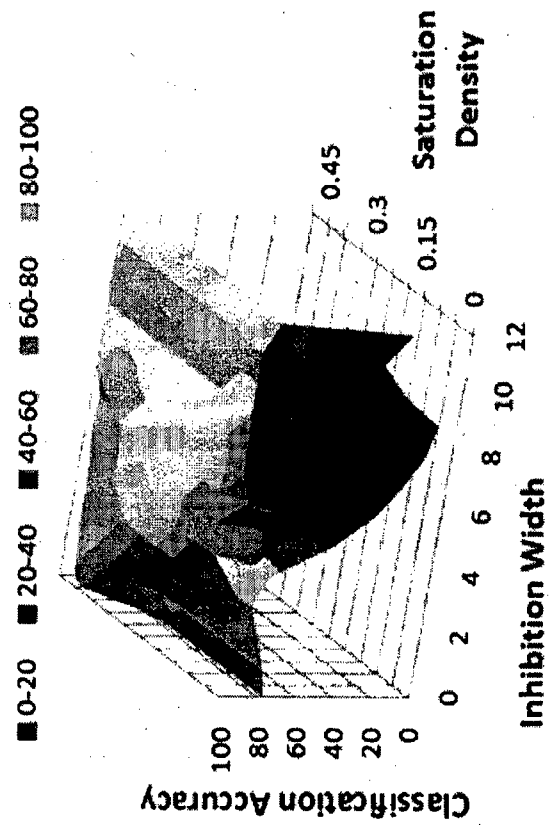


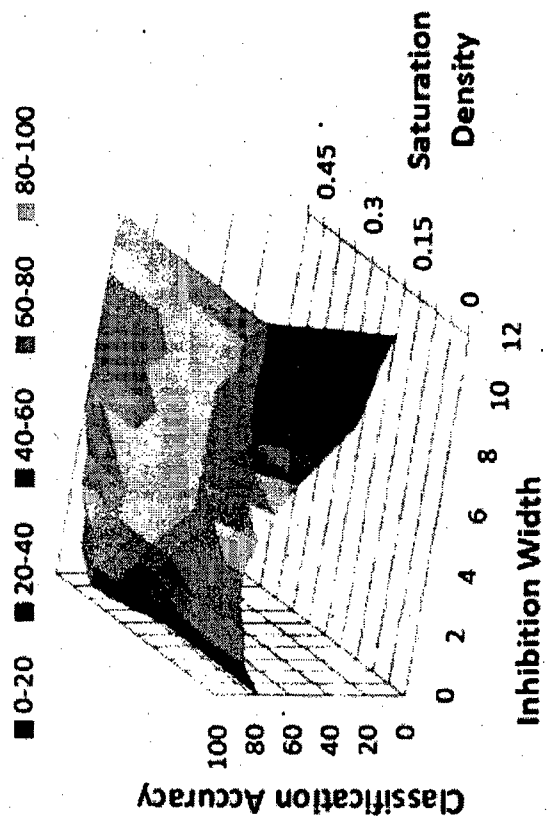
Figure 7

8/21



(b) 300 Channels and 0.055 Inhibition Strength

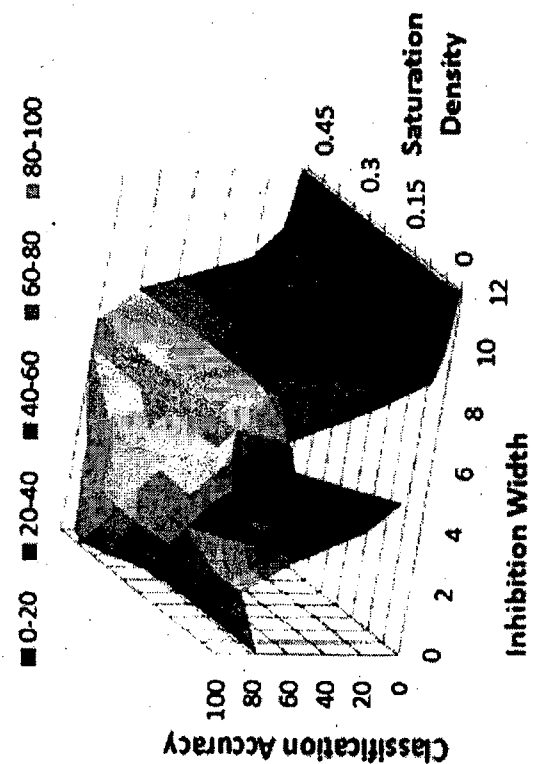
Figure 8B



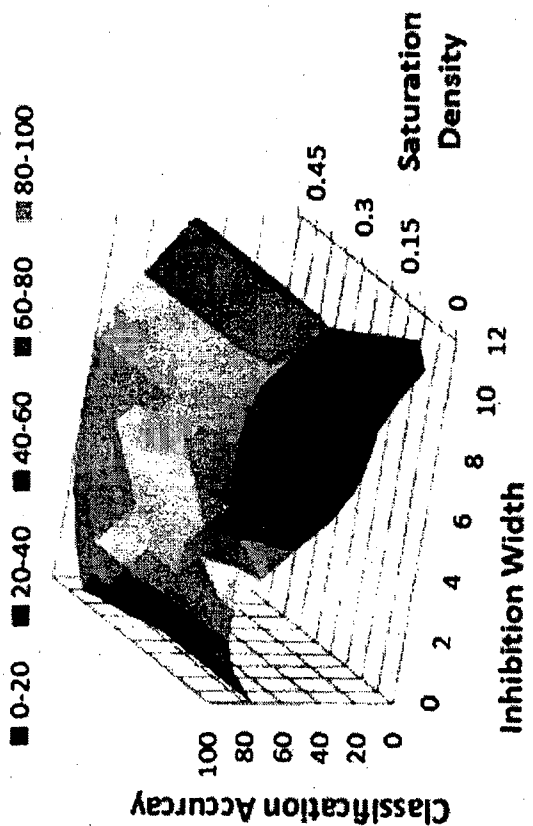
(a) 50 Channels and 0.055 Inhibition Strength

Figure 8A

9/21



(d) 300 Channels and 0.075 Inhibition Strength



(c) 50 Channels and 0.075 Inhibition Strength

Figure 8D

Figure 8C

10/21

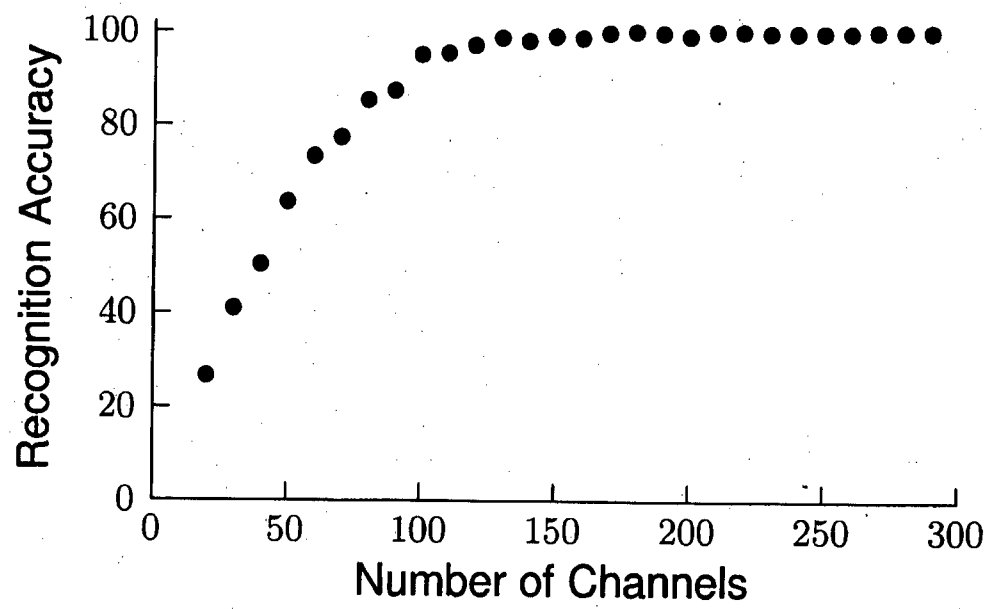


Figure 9

11/21

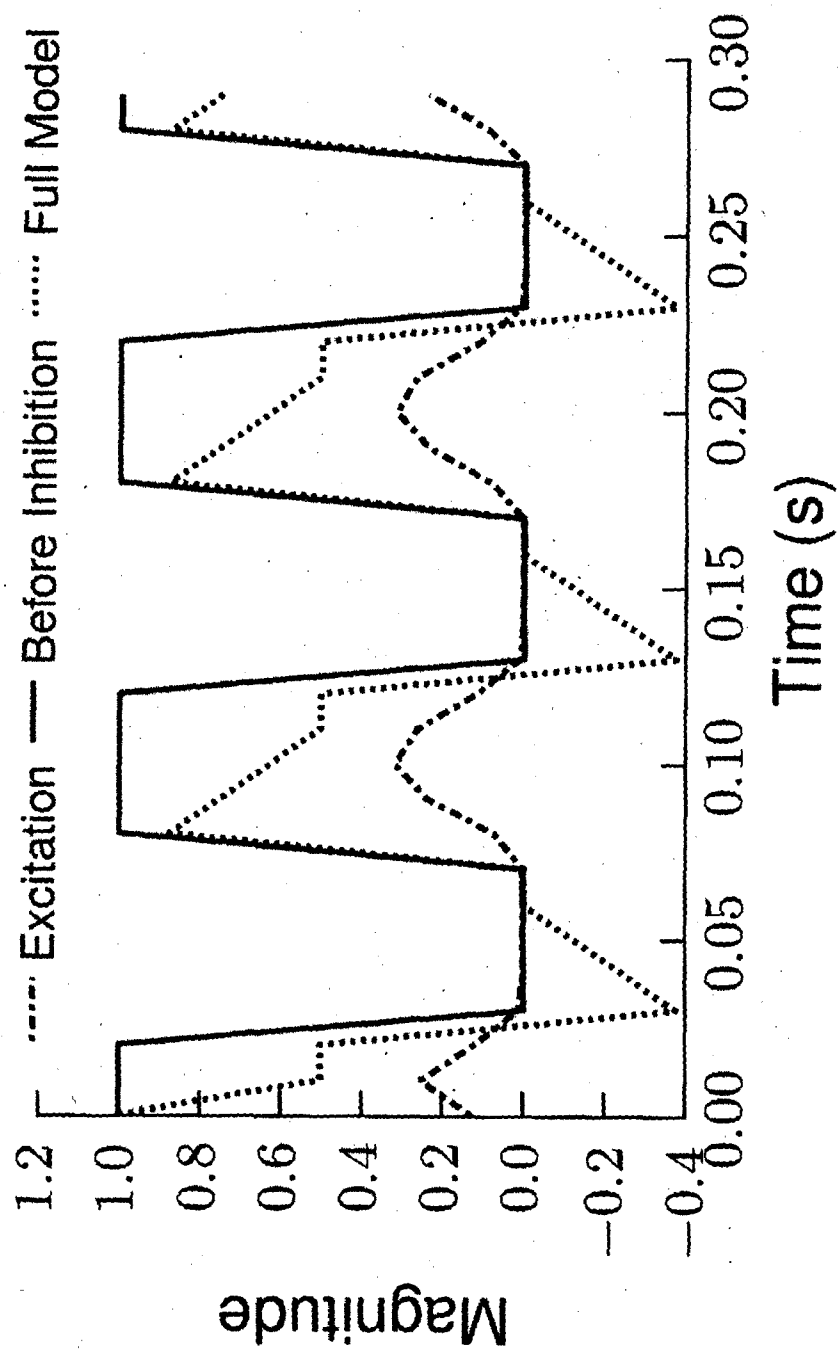


Figure 10

12/21

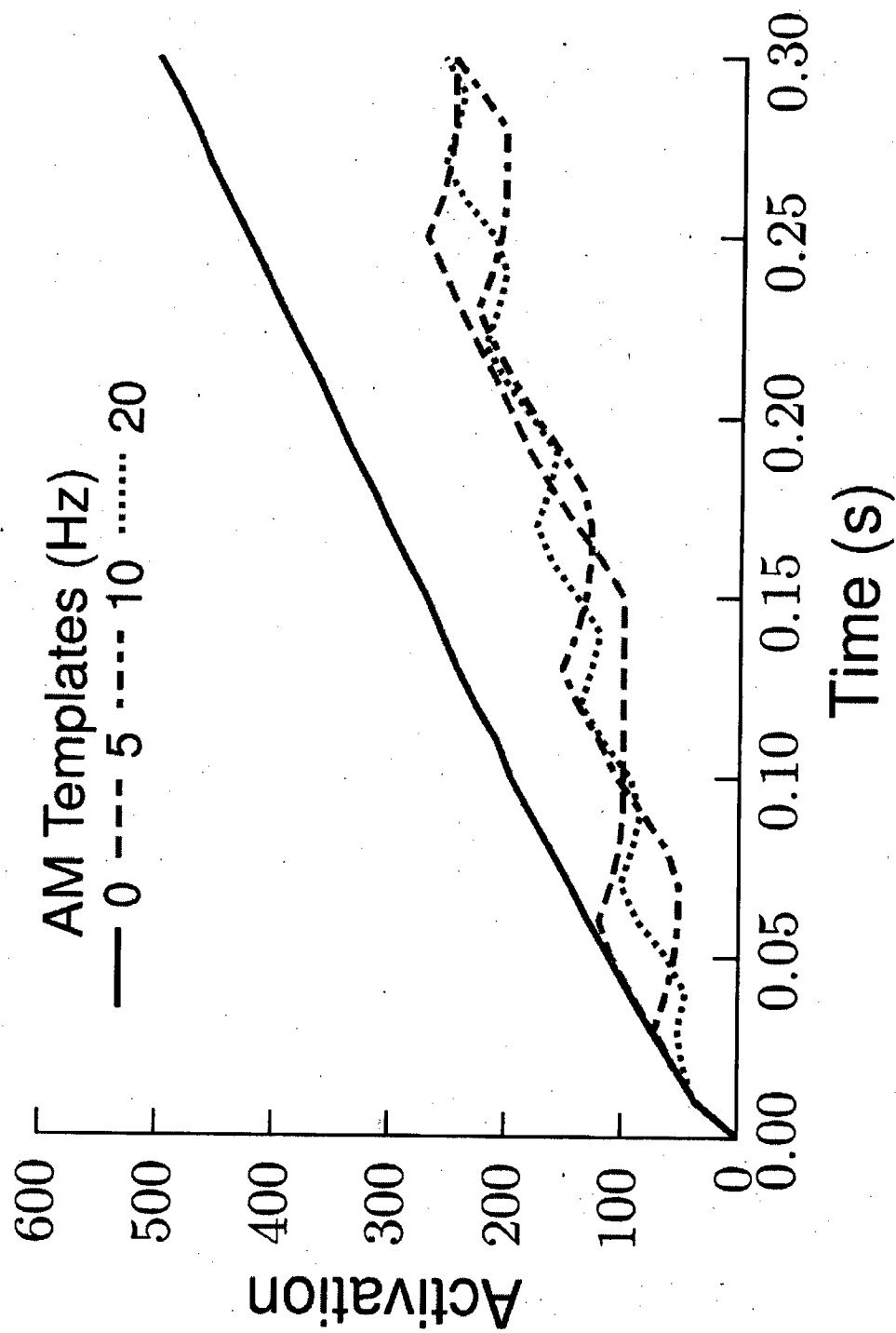


Figure 11

13/21

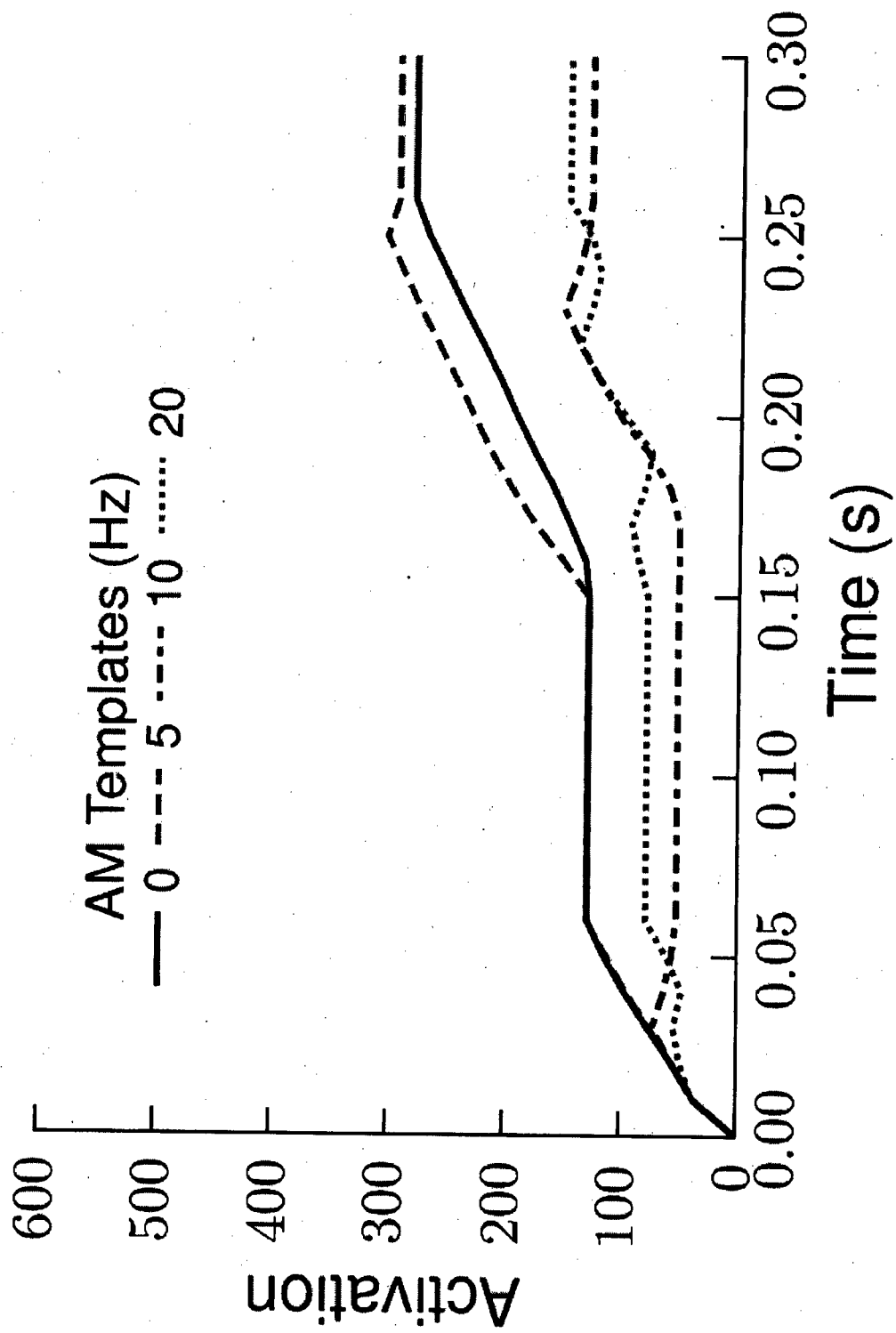


Figure 12

14/21

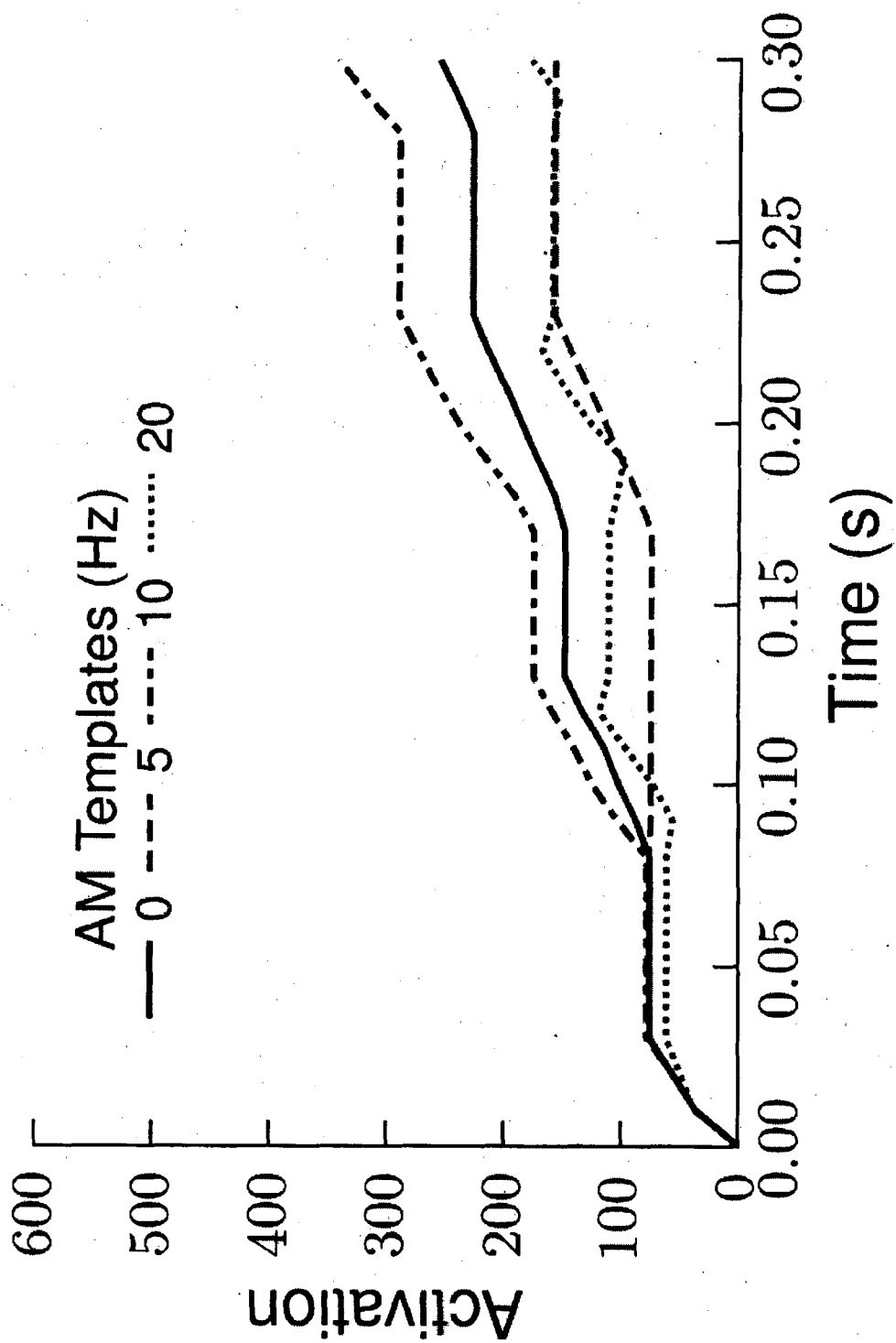


Figure 13

15/21

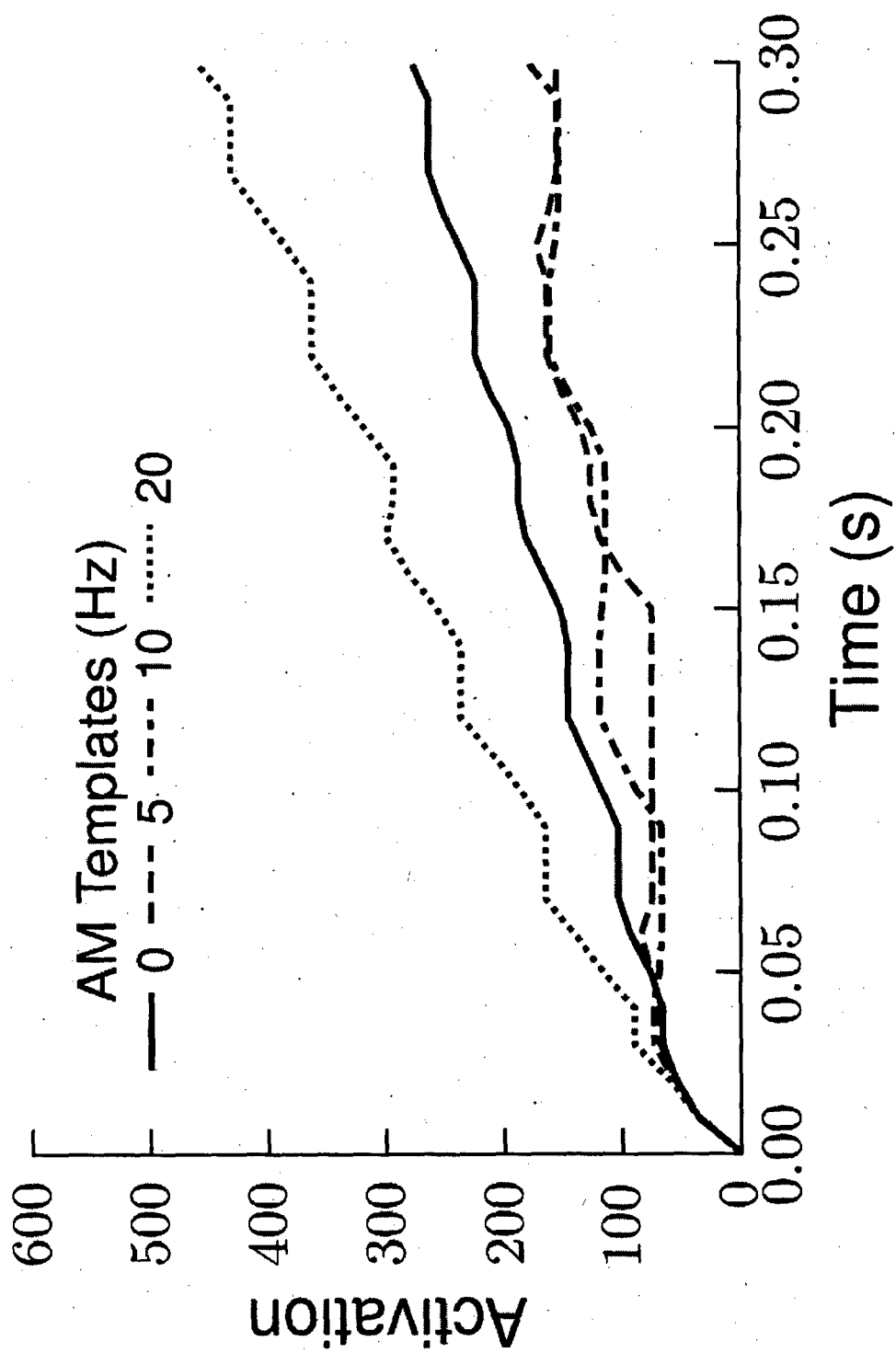


Figure 14

16/21

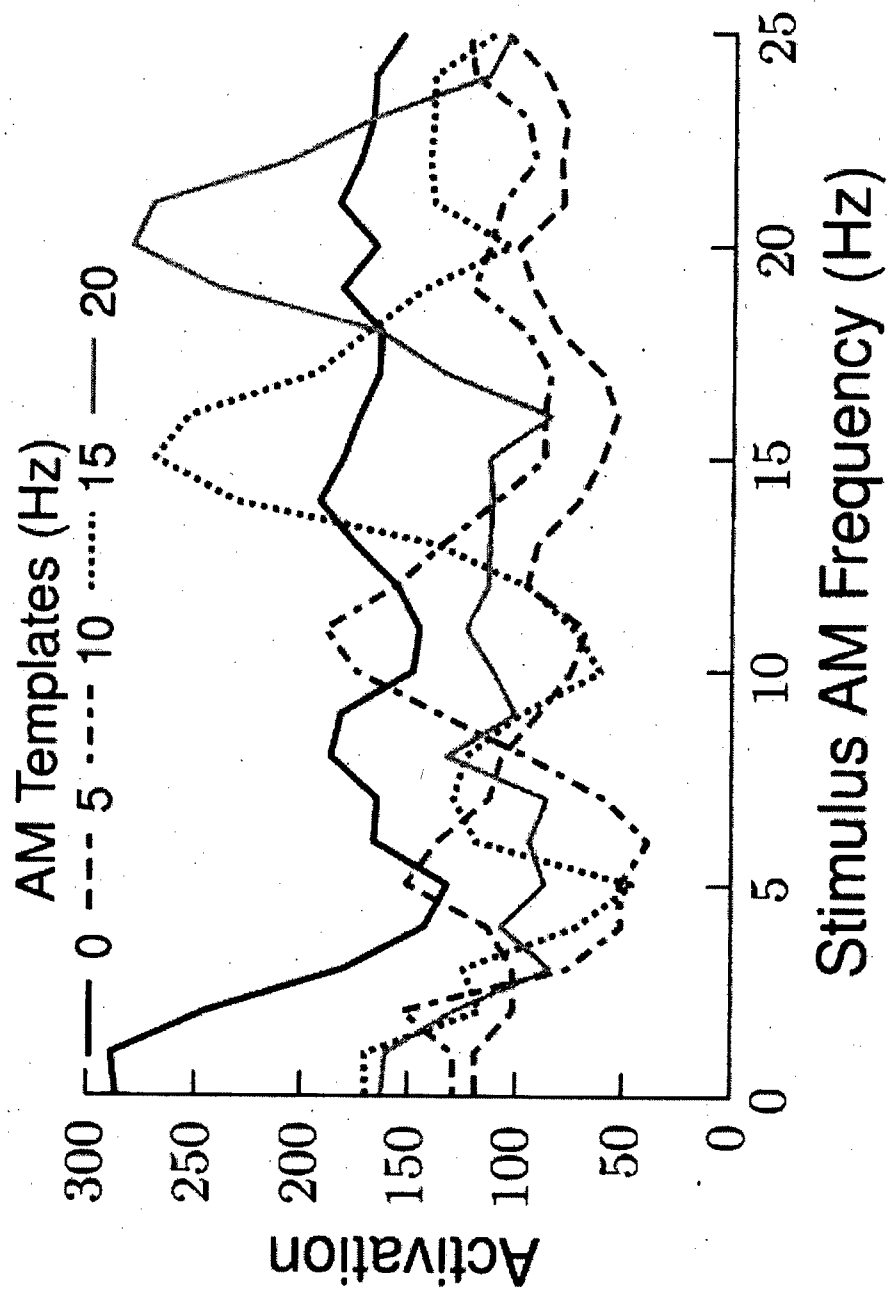


Figure 15

17/21

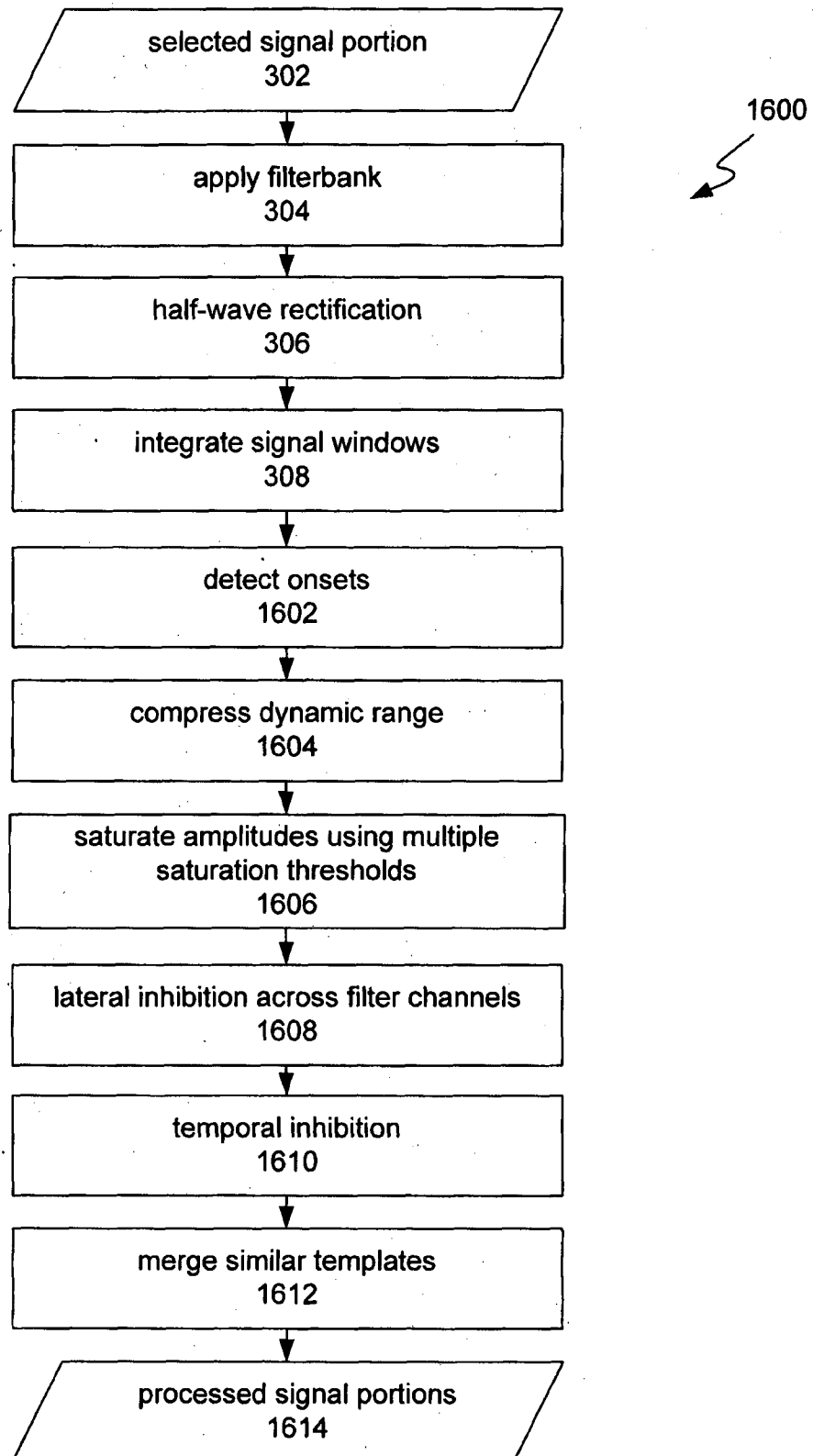


Figure 16

18/21

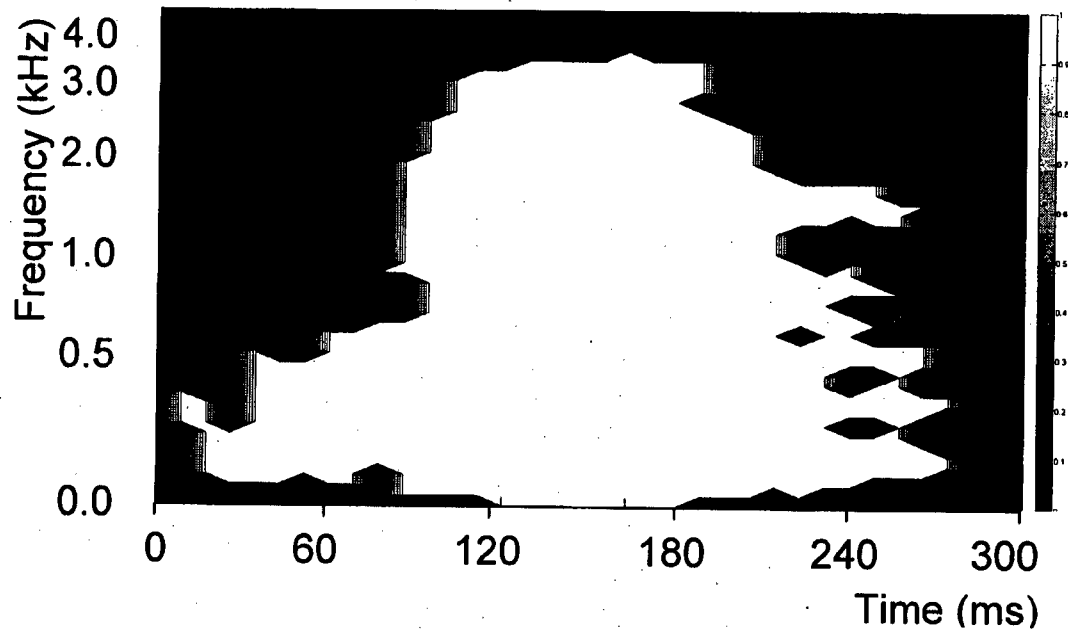


Figure 17A

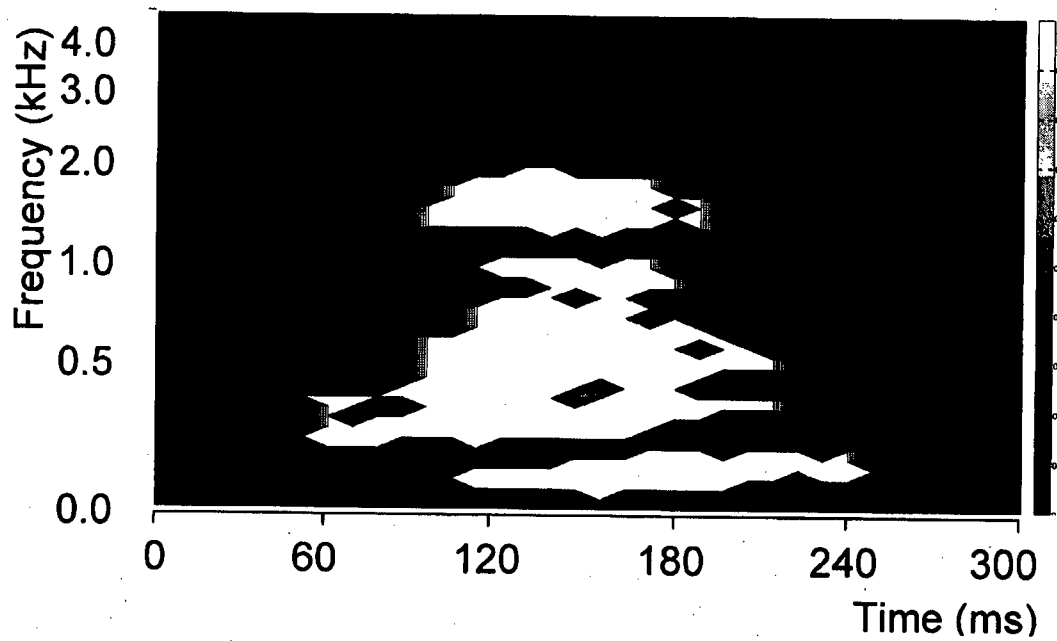


Figure 17B

19/21

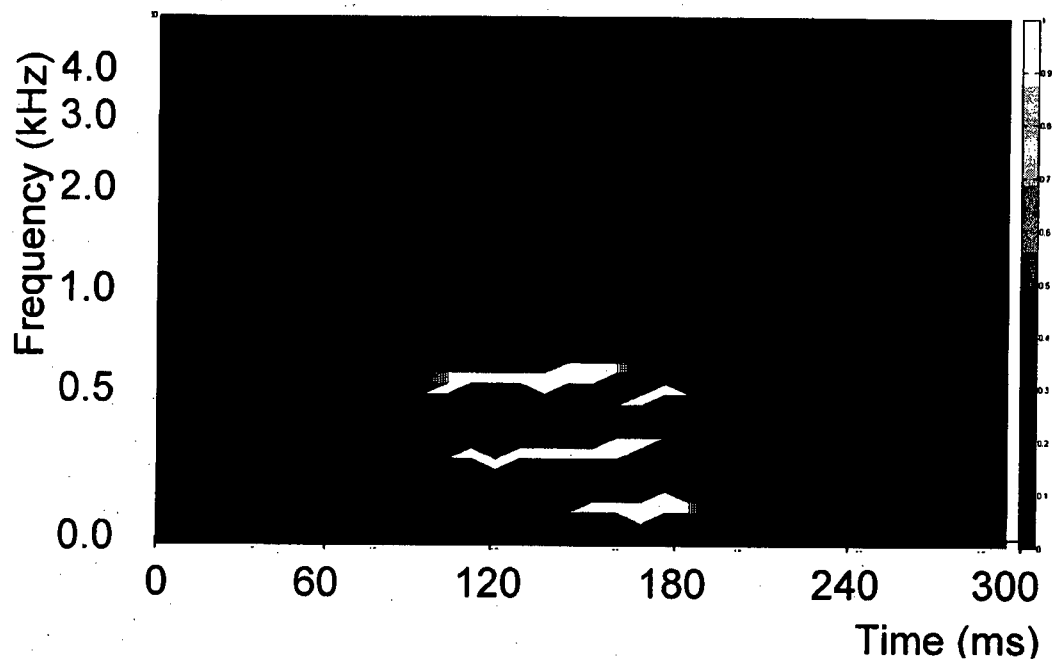


Figure 17C

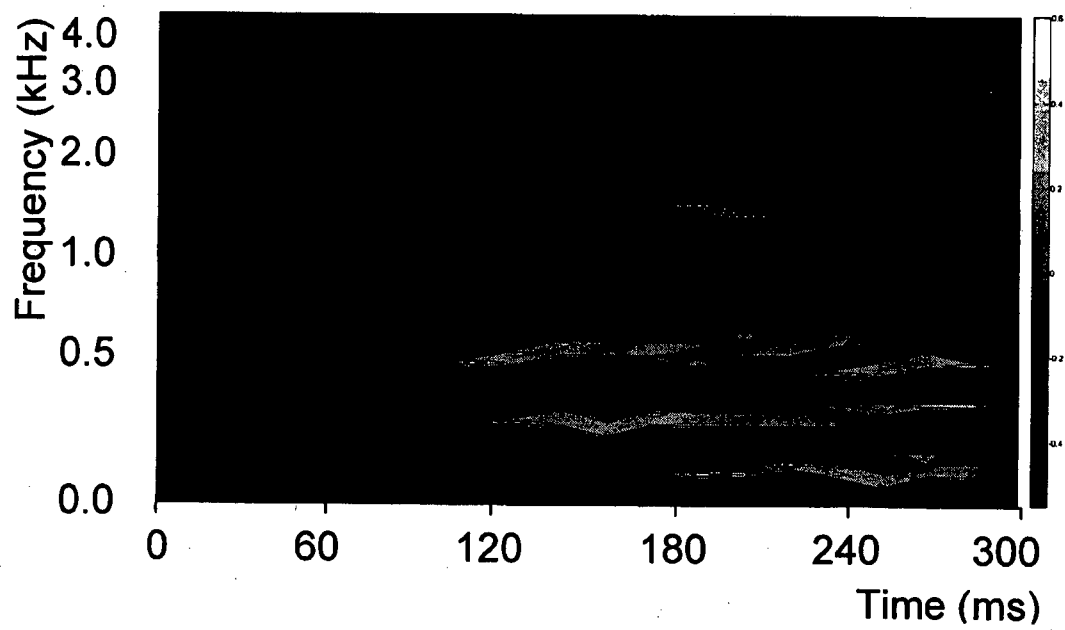


Figure 17D

20/21

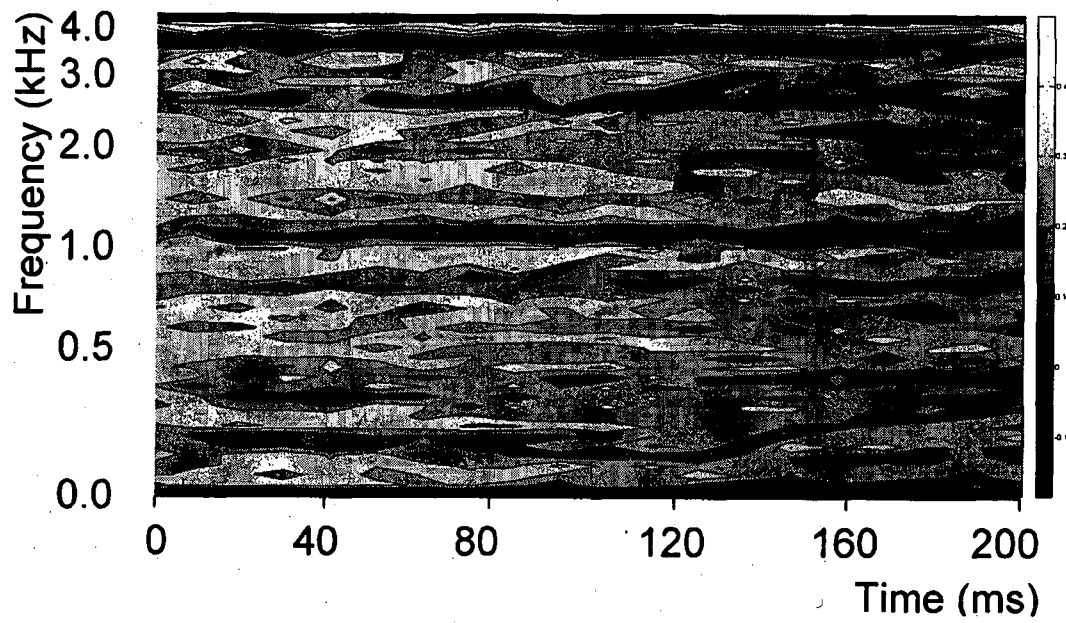


Figure 18

21/21

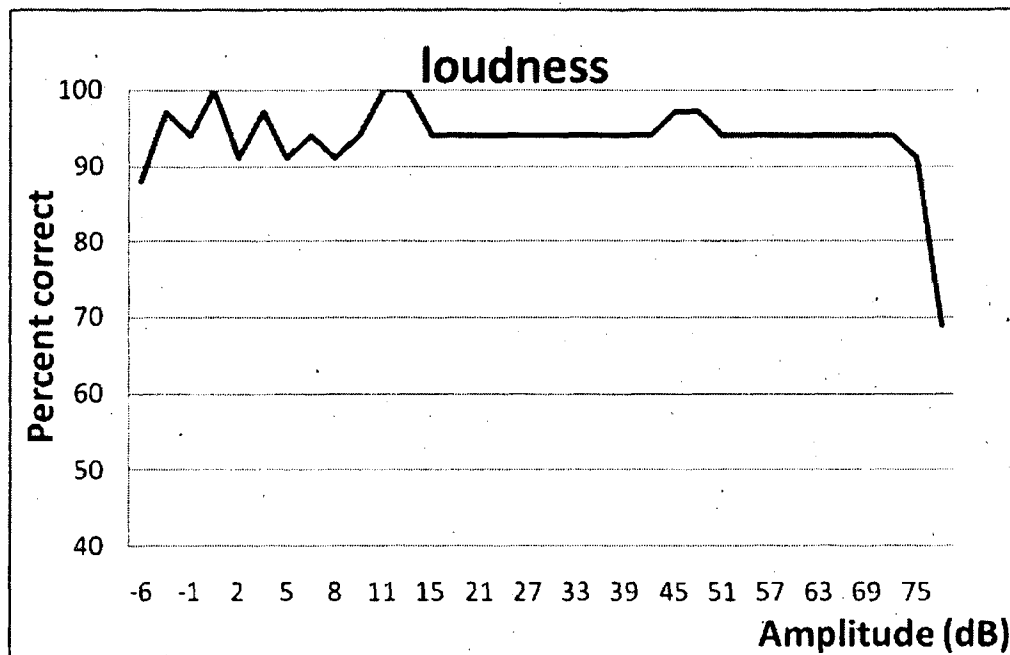


Figure 19

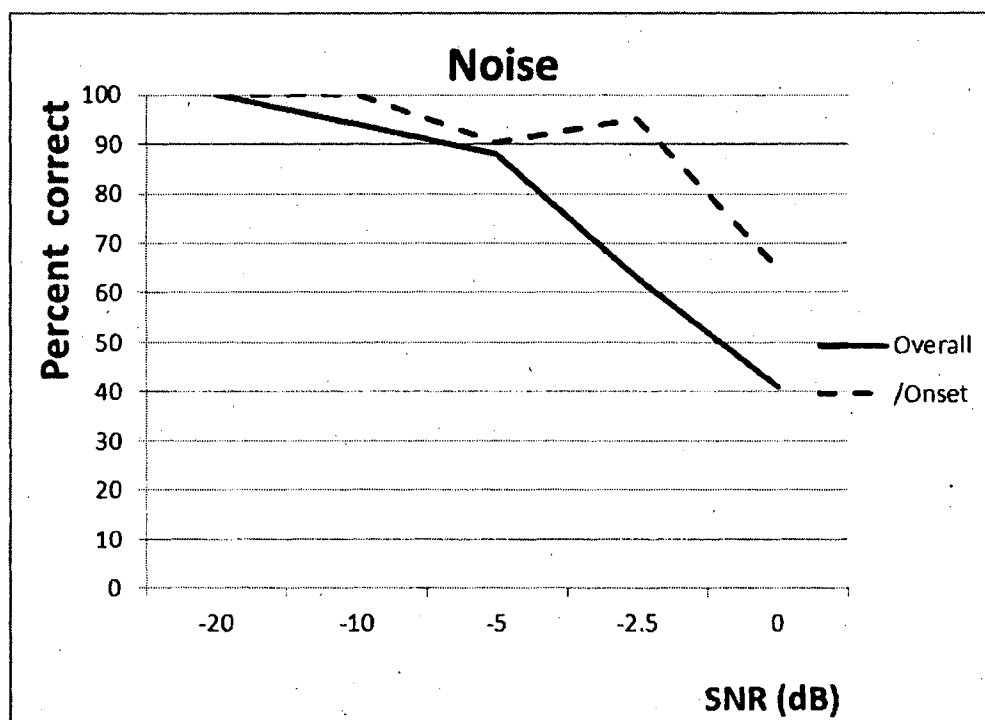


Figure 20

INTERNATIONAL SEARCH REPORT

International application No.
PCT/AU2012/001331

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/06 (2013.01)

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)

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)

WPI, EPODOC, Espacenet, Patent Lens, Google, IEEE Xplore, with keywords (audio, sound, signal, musical, speech, recognition, identification, classification, category, lateral inhibition, temporal inhibition, filter, set, generate, amplitude, frequency, training, learning, template, pattern, component, calculate, quantitative, score, measure, difference, correlation, relation, threshold, saturation, reinforcement) and the like terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Documents are listed in the continuation of Box C	



Further documents are listed in the continuation of Box C



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	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"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
"O" document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 4 January 2013	Date of mailing of the international search report 04 January 2013
Name and mailing address of the ISA/AU AUSTRALIAN PATENT OFFICE PO BOX 200, WODEN ACT 2606, AUSTRALIA Email address: pct@ipaustralia.gov.au Facsimile No.: +61 2 6283 7999	Authorised officer Ying Li AUSTRALIAN PATENT OFFICE (ISO 9001 Quality Certified Service) Telephone No. 0262832802

INTERNATIONAL SEARCH REPORT		International application No.
C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		PCT/AU2012/001331
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MCLACHLAN, N., 'A computational model of human pitch strength and height judgments', Hearing Research, 249 (1-2), March 2009, page 23-35.	1-3, 6, 16-24
Y	abstract, secs. [2, 3.1, 3.3, 3.4], pg. 32 col. 1, par. 3, pg. 34 col. 1 par. 2, fig. 8 secs. [2, 3.3], pg. 32 col. 1, par. 3, pg. 34 col. 1 par. 2, fig. 8	4-5, 10, 25
X	US 6947890 B1 (KITAZOE et al.) 20 September 2005	13-14
Y	col. 3 par. 2, col. 7 par. 2, claims [1, 10] col. 3 par. 2, col. 7 par. 2, claims [1, 10]	4-5, 10, 25

Form PCT/ISA/210 (fifth sheet) (July 2009)

