

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
30 April 2009 (30.04.2009)

PCT

(10) International Publication Number
WO 2009/055192 A1

(51) International Patent Classification:
G10L 19/14 (2006.01)

(74) Agents: **HAAS, Kenneth A.**, et al.; 1303 East Algonquin Road, Schaumburg, Illinois 60196 (US).

(21) International Application Number:
PCT/US2008/077693

(22) International Filing Date:
25 September 2008 (25.09.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/982,566 25 October 2007 (25.10.2007) US
12/187,423 7 August 2008 (07.08.2008) US

(71) Applicant (for all designated States except US): **MO-TOROLA, INC.** [US/US]; 1303 East Algonquin Road, Schaumburg, Illinois 60196 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **ASHLEY, James P.**, [US/US]; 1816 Arabian Avenue, Naperville, Illinois 60565 (US). **GIBBS, Jonathan A.**, [GB/GB]; 60, Church Street, Micheldever, Winchester Hampshire SO21 3DB (GB). **MITTAL, Udar**, [IN/US]; 475 Olive Street, Hoffman Estates, Illinois 60194 (US).

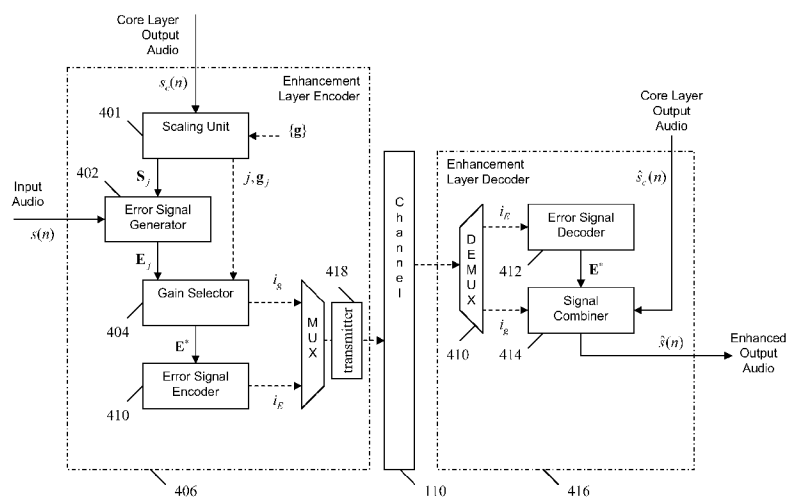
(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report
— with amended claims

(54) Title: METHOD AND APPARATUS FOR GENERATING AN ENHANCEMENT LAYER WITHIN AN AUDIO CODING SYSTEM



400

FIG. 4

(57) Abstract: During operation an input signal to be coded is received and coded to produce a coded audio signal. The coded audio signal is then scaled with a plurality of gain values to produce a plurality of scaled coded audio signals, each having an associated gain value and a plurality of error values are determined existing between the input signal and each of the plurality of scaled coded audio signals. A gain value is then chosen that is associated with a scaled coded audio signal resulting in a low error value existing between the input signal and the scaled coded audio signal. Finally, the low error value is transmitted along with the gain value as part of an enhancement layer to the coded audio signal.

WO 2009/055192 A1

METHOD AND APPARATUS FOR GENERATING AN ENHANCEMENT LAYER WITHIN AN AUDIO CODING SYSTEM

5 Field of the Invention

The present invention relates, in general, to communication systems and, more particularly, to coding speech and audio signals in such communication systems.

10 Background of the Invention

Compression of digital speech and audio signals is well known. Compression is generally required to efficiently transmit signals over a communications channel, or to store compressed signals on a digital media device, such as a solid-state memory device or computer hard disk. Although there are many compression (or “coding”) techniques, one method that has remained very popular for digital speech coding is known as Code Excited Linear Prediction (CELP), which is one of a family of “analysis-by-synthesis” coding algorithms. Analysis-by-synthesis generally refers to a coding process by which multiple parameters of a digital model are used to synthesize a set of candidate signals that are compared to an input signal and analyzed for distortion. A set of parameters that yield the lowest distortion is then either transmitted or stored, and eventually used to reconstruct an estimate of the original input signal. CELP is a particular analysis-by-synthesis method that uses one or more codebooks that each essentially comprises sets of code-vectors that are retrieved from the codebook in response to a codebook index.

In modern CELP coders, there is a problem with maintaining high quality speech and audio reproduction at reasonably low data rates. This is especially true for music or other generic audio signals that do not fit the CELP speech model very well. In this case, the model mismatch can cause severely degraded audio quality that can be unacceptable to an end user of the equipment that employs such methods. Therefore, there remains a need for improving performance of CELP type speech coders at low bit rates, especially for music and other non-speech type inputs.

35

Brief Description of the Drawings

FIG. 1 is a block diagram of a prior art embedded speech/audio compression system.

FIG. 2 is a more detailed example of the prior art enhancement layer encoder of FIG. 1.

FIG. 3 is a more detailed example of the prior art enhancement layer encoder of FIG. 1.

FIG. 4 is a block diagram of an enhancement layer encoder and decoder.

FIG. 5 is a block diagram of a multi-layer embedded coding system.

FIG. 6 is a block diagram of layer-4 encoder and decoder.

FIG. 7 is a flow chart showing operation of the encoders of FIG. 4 and FIG. 6.

15 Detailed Description of the Drawings

In order to address the above-mentioned need, a method and apparatus for generating an enhancement layer within an audio coding system is described herein. During operation an input signal to be coded is received and coded to produce a coded audio signal. The coded audio signal is then scaled with a plurality of gain values to produce a plurality of scaled coded audio signals, each having an associated gain value and a plurality of error values are determined existing between the input signal and each of the plurality of scaled coded audio signals. A gain value is then chosen that is associated with a scaled coded audio signal resulting in a low error value existing between the input signal and the scaled coded audio signal. Finally, the low error value is transmitted along with the gain value as part of an enhancement layer to the coded audio signal.

A prior art embedded speech/audio compression system is shown in FIG.1. The input audio $s(n)$ is first processed by a core layer encoder 102, which for these purposes may be a CELP type speech coding algorithm. The encoded bit-stream is transmitted to channel 110, as well as being input to a local core layer decoder 104, where the reconstructed core audio signal $s_c(n)$ is generated. The enhancement layer encoder 106 is then used to code additional information based on some comparison of signals $s(n)$ and $s_c(n)$, and may optionally use parameters from the core layer decoder 104. As in core layer decoder 104, core layer decoder 114 converts core layer bit-stream parameters to a core layer audio signal $\hat{s}_c(n)$. The enhancement layer decoder

116 then uses the enhancement layer bit-stream from channel 110 and signal $\hat{s}_c(n)$ to produce the enhanced audio output signal $\hat{s}(n)$.

The primary advantage of such an embedded coding system is that a particular channel 110 may not be capable of consistently supporting the bandwidth requirement associated with high quality audio coding algorithms. An embedded coder, however, allows a partial bit-stream to be received (e.g., only the core layer bit-stream) from the channel 110 to produce, for example, only the core output audio when the enhancement layer bit-stream is lost or corrupted. However, there are tradeoffs in quality between embedded vs. non-embedded coders, and also between different embedded coding optimization objectives. That is, higher quality enhancement layer coding can help achieve a better balance between core and enhancement layers, and also reduce overall data rate for better transmission characteristics (e.g., reduced congestion), which may result in lower packet error rates for the enhancement layers.

A more detailed example of a prior art enhancement layer encoder 106 is given in FIG. 2. Here, the error signal generator 202 is comprised of a weighted difference signal that is transformed into the MDCT (Modified Discrete Cosine Transform) domain for processing by error signal encoder 204. The error signal $\mathbf{E}$ is given as:

$$\mathbf{E} = \text{MDCT}\{\mathbf{W}(\mathbf{s} - \mathbf{s}_c)\}, \quad (1)$$

where $\mathbf{W}$ is a perceptual weighting matrix based on the LP (Linear Prediction) filter coefficients $A(z)$ from the core layer decoder 104, $\mathbf{s}$ is a vector (i.e., a *frame*) of samples from the input audio signal $s(n)$, and $\mathbf{s}_c$ is the corresponding vector of samples from the core layer decoder 104. An example MDCT process is described in ITU-T Recommendation G.729.1. The error signal $\mathbf{E}$ is then processed by the error signal encoder 204 to produce codeword i_E , which is subsequently transmitted to channel 110. For this example, it is important to note that error signal encoder 106 is presented with only one error signal $\mathbf{E}$ and outputs one associated codeword i_E . The reason for this will become apparent later.

The enhancement layer decoder 116 then receives the encoded bit-stream from channel 110 and appropriately de-multiplexes the bit-stream to produce codeword i_E . The error signal decoder 212 uses codeword i_E to reconstruct the enhancement layer error signal $\hat{\mathbf{E}}$, which is then combined with the core layer output audio signal $\hat{s}_c(n)$ as follows, to produce the enhanced audio output signal $\hat{s}(n)$:

$$\hat{\mathbf{s}} = \mathbf{s}_c + \mathbf{W}^{-1} \text{MDCT}^{-1} \{ \hat{\mathbf{E}} \}, \quad (2)$$

where MDCT^{-1} is the inverse MDCT (including overlap-add), and $\mathbf{W}^{-1}$ is the inverse perceptual weighting matrix.

5 Another example of an enhancement layer encoder is shown in FIG. 3. Here, the generation of the error signal $\mathbf{E}$ by error signal generator 302 involves adaptive pre-scaling, in which some modification to the core layer audio output $s_c(n)$ is performed. This process results in some number of bits to be generated, which are shown in enhancement layer encoder 106 as codeword i_s .

10 Additionally, enhancement layer encoder 106 shows the input audio signal $s(n)$ and transformed core layer output audio $\mathbf{S}_c$ being inputted to error signal encoder 304. These signals are used to construct a psychoacoustic model for improved coding of the enhancement layer error signal $\mathbf{E}$. Codewords i_s and i_E are then multiplexed by MUX 308, and then sent to channel 110 for subsequent decoding by enhancement
15 layer decoder 116. The coded bit-stream is received by demux 310, which separates the bit-stream into components i_s and i_E . Codeword i_E is then used by error signal decoder 312 to reconstruct the enhancement layer error signal $\hat{\mathbf{E}}$. Signal combiner 314 scales signal $\hat{s}_c(n)$ in some manner using scaling bits i_s , and then combines the result with the enhancement layer error signal $\hat{\mathbf{E}}$ to produce the enhanced audio
20 output signal $\hat{s}(n)$.

A first embodiment of the present invention is given in FIG. 4. This figure shows enhancement layer encoder 406 receiving core layer output signal $s_c(n)$ by scaling unit 401. A predetermined set of gains $\{\mathbf{g}\}$ is used to produce a plurality of scaled core layer output signals $\{\mathbf{S}\}$, where $\mathbf{g}_j$ and $\mathbf{S}_j$ are the j -th candidates of the
25 respective sets. Within scaling unit 401, the first embodiment processes signal $s_c(n)$ in the (MDCT) domain as:

$$\mathbf{S}_j = \mathbf{G}_j \times \text{MDCT} \{ \mathbf{W} \mathbf{s}_c \}, \quad 0 \leq j < M, \quad (3)$$

30 where $\mathbf{W}$ may be some perceptual weighting matrix, $\mathbf{s}_c$ is a vector of samples from the core layer decoder 104, the MDCT is an operation well known in the art, and $\mathbf{G}_j$ may be a gain matrix formed by utilizing a gain vector candidate $\mathbf{g}_j$, and where M is the number gain vector candidates. In the first embodiment, $\mathbf{G}_j$ uses vector $\mathbf{g}_j$ as the diagonal and zeros everywhere else (i.e., a diagonal matrix), although many
35 possibilities exist. For example, $\mathbf{G}_j$ may be a band matrix, or may even be a simple scalar quantity multiplied by the identity matrix $\mathbf{I}$. Alternatively, there may be some

5 advantage to leaving the signal S_j in the time domain or there may be cases where it is advantageous to transform the audio to a different domain, such as the Discrete Fourier Transform (DFT) domain. Many such transforms are well known in the art. In these cases, the scaling unit may output the appropriate S_j based on the respective vector domain.

But in any case, the primary reason to scale the core layer output audio is to compensate for model mismatch (or some other coding deficiency) that may cause significant differences between the input signal and the core layer codec. For example, if the input audio signal is primarily a music signal and the core layer codec is based on a speech model, then the core layer output may contain severely distorted signal characteristics, in which case, it is beneficial from a sound quality perspective to selectively reduce the energy of this signal component prior to applying supplemental coding of the signal by way of one or more enhancement layers.

The gain scaled core layer audio candidate vector S_j and input audio $s(n)$ may then be used as input to error signal generator 402. In the preferred embodiment of the present invention, the input audio signal $s(n)$ is converted to vector S such that S and S_j are correspondingly aligned. That is, the vector s representing $s(n)$ is time (phase) aligned with s_c , and the corresponding operations may be applied so that in the preferred embodiment:

$$20 \quad E_j = \text{MDCT}\{\mathbf{W}\mathbf{s}\} - S_j; \quad 0 \leq j < M. \quad (4)$$

This expression yields a plurality of error signal vectors E_j that represent the weighted difference between the input audio and the gain scaled core layer output audio in the MDCT spectral domain. In other embodiments where different domains are considered, the above expression may be modified based on the respective processing domain.

Gain selector 404 is then used to evaluate the plurality of error signal vectors E_j , in accordance with the first embodiment of the present invention, to produce an optimal error vector E^* , an optimal gain parameter g^* , and subsequently, a corresponding gain index i_g . The gain selector 404 may use a variety of methods to determine the optimal parameters, E^* and g^* , which may involve closed loop methods (e.g., minimization of a distortion metric), open loop methods (e.g., heuristic classification, model performance estimation, etc.), or a combination of both methods.

35 In the preferred embodiment, a biased distortion metric may be used, which is given as the biased energy difference between the original audio signal vector S and the composite reconstructed signal vector:

$$j^* = \arg \min_{0 \leq j < M} \left\{ \beta_j \cdot \left\| \mathbf{S} - (\mathbf{S}_j + \hat{\mathbf{E}}_j) \right\|^2 \right\}, \quad (5)$$

where $\hat{\mathbf{E}}_j$ may be the quantified estimate of the error signal vector $\mathbf{E}_j$, and β_j may be a
 5 bias term which is used to supplement the decision of choosing the perceptually
 optimal gain error index j^* . An exemplary method for vector quantization of a signal
 vector is given in US Patent Application Serial No. 11/531122, entitled APPARATUS
 AND METHOD FOR LOW COMPLEXITY COMBINATORIAL CODING OF SIGNALS,
 although many other methods are possible. Recognizing that $\mathbf{E}_j = \mathbf{S} - \mathbf{S}_j$, equation (5)
 10 may be rewritten as:

$$j^* = \arg \min_{0 \leq j < M} \left\{ \beta_j \cdot \left\| \mathbf{E}_j - \hat{\mathbf{E}}_j \right\|^2 \right\}. \quad (6)$$

In this expression, the term $\varepsilon_j = \left\| \mathbf{E}_j - \hat{\mathbf{E}}_j \right\|^2$ represents the energy of the difference
 15 between the unquantized and quantized error signals. For clarity, this quantity may be
 referred to as the “residual energy”, and may further be used to evaluate a “gain
 selection criterion”, in which the optimum gain parameter $\mathbf{g}^*$ is selected. One such
 gain selection criterion is given in equation (6), although many are possible.

The need for a bias term β_j may arise from the case where the error weighting
 20 function $\mathbf{W}$ in equations (3) and (4) may not adequately produce equally perceptible
 distortions across vector $\hat{\mathbf{E}}_j$. For example, although the error weighting function $\mathbf{W}$
 may be used to attempt to “whiten” the error spectrum to some degree, there may be
 certain advantages to placing more weight on the low frequencies, due to the
 perception of distortion by the human ear. As a result of increased error weighting in
 25 the low frequencies, the high frequency signals may be under-modeled by the
 enhancement layer. In these cases, there may be a direct benefit to biasing the
 distortion metric towards values of $\mathbf{g}_j$ that do not attenuate the high frequency
 components of $\mathbf{S}_j$, such that the under-modeling of high frequencies does not result in
 objectionable or unnatural sounding artifacts in the final reconstructed audio signal.
 30 One such example would be the case of an unvoiced speech signal. In this case, the
 input audio is generally made up of mid to high frequency noise-like signals produced
 from turbulent flow of air from the human mouth. It may be that the core layer
 encoder does not code this type of waveform directly, but may use a noise model to
 generate a similar sounding audio signal. This may result in a generally low

correlation between the input audio and the core layer output audio signals. However, in this embodiment, the error signal vector $\mathbf{E}_j$ is based on a difference between the input audio and core layer audio output signals. Since these signals may not be correlated very well, the energy of the error signal $\mathbf{E}_j$ may not necessarily be lower than either the input audio or the core layer output audio. In that case, minimization of the error in equation (6) may result in the gain scaling being too aggressive, which may result in potential audible artifacts.

In another case, the bias factors β_j may be based on other signal characteristics of the input audio and/or core layer output audio signals. For example, the peak-to-average ratio of the spectrum of a signal may give an indication of that signal's harmonic content. Signals such as speech and certain types of music may have a high harmonic content and thus a high peak-to-average ratio. However, a music signal processed through a speech codec may result in a poor quality due to coding model mismatch, and as a result, the core layer output signal spectrum may have a reduced peak-to-average ratio when compared to the input signal spectrum. In this case, it may be beneficial reduce the amount of bias in the minimization process in order to allow the core layer output audio to be gain scaled to a lower energy thereby allowing the enhancement layer coding to have a more pronounced effect on the composite output audio. Conversely, certain types speech or music input signals may exhibit lower peak-to-average ratios, in which case, the signals may be perceived as being more noisy, and may therefore benefit from less scaling of the core layer output audio by increasing the error bias. An example of a function to generate the bias factors for β_j , is given as:

$$\beta_j = \begin{cases} 1 + 10^6 \cdot j; & UVSpeech == TRUE \text{ or } \phi_s < \lambda \phi_{s_c}, \\ 10^{(-j \cdot \Delta/10)}; & \text{otherwise} \end{cases}, \quad 0 \leq j < M. \quad (7)$$

where λ may be some threshold, and the peak-to-average ratio for vector ϕ_y may be given as:

$$\phi_y = \frac{\max \{ |\mathbf{y}_{k_1 k_2}| \}}{\frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} |y(k)|}, \quad (8)$$

and where $\mathbf{y}_{k_1 k_2}$ is a vector subset of $y(k)$ such that $\mathbf{y}_{k_1 k_2} = y(k); k_1 \leq k \leq k_2$.

Once the optimum gain index j^* is determined from equation (6), the associated codeword i_g is generated and the optimum error vector $\mathbf{E}^*$ is sent to error

signal encoder 410, where $\mathbf{E}^*$ is coded into a form that is suitable for multiplexing with other codewords (by MUX 408) and transmitted for use by a corresponding decoder. In the preferred embodiment, error signal encoder 408 uses Factorial Pulse Coding (FPC). This method is advantageous from a processing complexity point of view since the enumeration process associated with the coding of vector $\mathbf{E}^*$ is independent of the vector generation process that is used to generate $\hat{\mathbf{E}}_j$.

Enhancement layer decoder 416 reverses these processes to produce the enhance audio output $\hat{s}(n)$. More specifically, i_g and i_E are received by decoder 416, with i_E being sent to error signal decoder 412 where the optimum error vector $\mathbf{E}^*$ is derived from the codeword. The optimum error vector $\mathbf{E}^*$ is passed to signal combiner 414 where the received $\hat{s}_c(n)$ is modified as in equation (2) to produce $\hat{s}(n)$.

A second embodiment of the present invention involves a multi-layer embedded coding system as shown in FIG. 5. Here, it can be seen that there are five embedded layers given for this example. Layers 1 and 2 may be both speech codec based, and layers 3, 4, and 5 may be MDCT enhancement layers. Thus, encoders 502 and 503 may utilize speech codecs to produce and output encoded input signal $s(n)$. Encoders 510, 512, and 514 comprise enhancement layer encoders, each outputting a differing enhancement to the encoded signal. Similar to the previous embodiment, the error signal vector for layer 3 (encoder 510) may be given as:

$$\mathbf{E}_3 = \mathbf{S} - \mathbf{S}_2, \quad (9)$$

where $\mathbf{S} = \text{MDCT}\{\mathbf{W}\mathbf{s}\}$ is the weighted transformed input signal, and $\mathbf{S}_2 = \text{MDCT}\{\mathbf{W}\mathbf{s}_2\}$ is the weighted transformed signal generated from the layer 1/2 decoder 506. In this embodiment, layer 3 may be a low rate quantization layer, and as such, there may be relatively few bits for coding the corresponding quantized error signal $\hat{\mathbf{E}}_3 = Q\{\mathbf{E}_3\}$. In order to provide good quality under these constraints, only a fraction of the coefficients within $\mathbf{E}_3$ may be quantized. The positions of the coefficients to be coded may be fixed or may be variable, but if allowed to vary, it may be required to send additional information to the decoder to identify these positions. If, for example, the range of coded positions starts at k_s and ends at k_e , where $0 \leq k_s < k_e < N$, then the quantized error signal vector $\hat{\mathbf{E}}_3$ may contain non-zero values only within that range, and zeros for positions outside that range. The position and range information may also be implicit, depending on the coding method used. For example, it is well known in audio coding that a band of frequencies may be deemed perceptually important, and that coding of a signal vector may focus on those frequencies. In these circumstances, the coded range may be variable, and may not

span a contiguous set of frequencies. But at any rate, once this signal is quantized, the composite coded output spectrum may be constructed as:

$$\mathbf{S}_3 = \hat{\mathbf{E}}_3 + \mathbf{S}_2, \quad (10)$$

5

which is then used as input to layer 4 encoder 512.

Layer 4 encoder 512 is similar to the enhancement layer encoder 406 of the previous embodiment. Using the gain vector candidate $\mathbf{g}_j$, the corresponding error vector may be described as:

10

$$\mathbf{E}_4(j) = \mathbf{S} - \mathbf{G}_j \mathbf{S}_3, \quad (11)$$

15

where $\mathbf{G}_j$ may be a gain matrix with vector $\mathbf{g}_j$ as the diagonal component. In the current embodiment, however, the gain vector $\mathbf{g}_j$ may be related to the quantized error signal vector $\hat{\mathbf{E}}_3$ in the following manner. Since the quantized error signal vector $\hat{\mathbf{E}}_3$ may be limited in frequency range, for example, starting at vector position k_s and ending at vector position k_e , the layer 3 output signal $\mathbf{S}_3$ is presumed to be coded fairly accurately within that range. Therefore, in accordance with the present invention, the gain vector $\mathbf{g}_j$ is adjusted based on the coded positions of the layer 3 error signal vector, k_s and k_e . More specifically, in order to preserve the signal integrity at those locations, the corresponding individual gain elements may be set to a constant value α . That is:

20

$$g_j(k) = \begin{cases} \alpha; & k_s \leq k \leq k_e \\ \gamma_j(k); & \text{otherwise} \end{cases}, \quad (12)$$

25

where generally $0 \leq \gamma_j(k) \leq 1$ and $g_j(k)$ is the gain of the k -th position of the j -th candidate vector. In the preferred embodiment, the value of the constant is one ($\alpha = 1$), however many values are possible. In addition, the frequency range may span multiple starting and ending positions. That is, equation (12) may be segmented into non-continuous ranges of varying gains that are based on some function of the error signal $\hat{\mathbf{E}}_3$, and may be written more generally as:

30

$$g_j(k) = \begin{cases} \alpha; & \hat{E}_3(k) \neq 0 \\ \gamma_j(k); & \text{otherwise} \end{cases}, \quad (13)$$

For this example, a fixed gain α is used to generate $g_j(k)$ when the corresponding positions in the previously quantized error signal $\hat{\mathbf{E}}_3$ are non-zero, and gain function $\gamma_j(k)$ is used when the corresponding positions in $\hat{\mathbf{E}}_3$ are zero. One possible gain function may be defined as:

5

$$\gamma_j(k) = \begin{cases} \alpha \cdot 10^{(-j \cdot \Delta / 20)}; & k_l \leq k \leq k_h, \\ \alpha; & \text{otherwise} \end{cases}, \quad 0 \leq j < M, \quad (14)$$

where Δ is a step size (e.g., $\Delta \approx 2.2$ dB), α is a constant, M is the number of candidates (e.g., $M = 4$, which can be represented using only 2 bits), and k_l and k_h are the low and high frequency cutoffs, respectively, over which the gain reduction may take place. The introduction of parameters k_l and k_h is useful in systems where scaling is desired only over a certain frequency range. For example, in a given embodiment, the high frequencies may not be adequately modeled by the core layer, thus the energy within the high frequency band may be inherently lower than that in the input audio signal. In that case, there may be little or no benefit from scaling the layer 3 output in that region signal since the overall error energy may increase as a result.

15

Summarizing, the plurality of gain vector candidates $\mathbf{g}_j$ is based on some function of the coded elements of a previously coded signal vector, in this case $\hat{\mathbf{E}}_3$. This can be expressed in general terms as:

20

$$g_j(k) = f(k, \hat{\mathbf{E}}_3). \quad (15)$$

The corresponding decoder operations are shown on the right hand side of FIG. 5. As the various layers of coded bit-streams (i_1 to i_5) are received, the higher quality output signals are built on the hierarchy of enhancement layers over the core layer (layer 1) decoder. That is, for this particular embodiment, as the first two layers are comprised of time domain speech model coding (e.g., CELP) and the remaining three layers are comprised of transform domain coding (e.g., MDCT), the final output for the system $\hat{s}(n)$ is generated according to the following:

30

$$\hat{s}(n) = \begin{cases} \hat{s}_1(n); \\ \hat{s}_2(n) = \hat{s}_1(n) + \hat{e}_2(n); \\ \hat{s}_3(n) = \mathbf{W}^{-1} \text{MDCT}^{-1} \left\{ \hat{\mathbf{S}}_2 + \hat{\mathbf{E}}_3 \right\}; \\ \hat{s}_4(n) = \mathbf{W}^{-1} \text{MDCT}^{-1} \left\{ \mathbf{G}_j \cdot \left(\hat{\mathbf{S}}_2 + \hat{\mathbf{E}}_3 \right) + \hat{\mathbf{E}}_4 \right\}; \\ \hat{s}_5(n) = \mathbf{W}^{-1} \text{MDCT}^{-1} \left\{ \mathbf{G}_j \cdot \left(\hat{\mathbf{S}}_2 + \hat{\mathbf{E}}_3 \right) + \hat{\mathbf{E}}_4 + \hat{\mathbf{E}}_5 \right\}; \end{cases}, \quad (16)$$

where $\hat{e}_2(n)$ is the layer 2 time domain enhancement layer signal, and $\hat{\mathbf{S}}_2 = \text{MDCT}\{\mathbf{W}\mathbf{s}_2\}$ is the weighted MDCT vector corresponding to the layer 2 audio output $\hat{s}_2(n)$. In this expression, the overall output signal $\hat{s}(n)$ may be determined
 5 from the highest level of consecutive bit-stream layers that are received. In this embodiment, it is assumed that lower level layers have a higher probability of being properly received from the channel, therefore, the codeword sets $\{i_1\}$, $\{i_1 i_2\}$, $\{i_1 i_2 i_3\}$, etc., determine the appropriate level of enhancement layer decoding in equation (16).

FIG. 6 is a block diagram showing layer 4 encoder 512 and decoder 522. The
 10 encoder and decoder shown in FIG. 6 are similar to those shown in FIG. 4, except that the gain value used by scaling units 601 and 618 is derived via frequency selective gain generators 603 and 616, respectively. During operation layer 3 audio output $\mathbf{S}_3$ is output from layer 3 encoder and received by scaling unit 601. Additionally, layer 3 error vector $\hat{\mathbf{E}}_3$ is output from layer 3 encoder 510 and received by frequency
 15 selective gain generator 603. As discussed, since the quantized error signal vector $\hat{\mathbf{E}}_3$ may be limited in frequency range, the gain vector $\mathbf{g}_j$ is adjusted based on, for example, the positions k_s and k_e as shown in equation 12, or the more general expression in equation 13.

The scaled audio $\mathbf{S}_j$ is output from scaling unit 601 and received by error
 20 signal generator 602. As discussed above, error signal generator 602 receives the input audio signal $\mathbf{S}$ and determines an error value $\mathbf{E}_j$ for each scaling vector utilized by scaling unit 601. These error vectors are passed to gain selector circuitry 604 along with the gain values used in determining the error vectors and a particular error $\mathbf{E}^*$ based on the optimal gain value $\mathbf{g}^*$. A codeword (i_g) representing the optimal gain $\mathbf{g}^*$
 25 is output from gain selector 604, along with the optimal error vector $\mathbf{E}^*$, is passed to encoder 610 where codeword i_E is determined and output. Both i_g and i_E are output to multiplexer 608 and transmitted via channel 110 to layer 4 decoder 522.

During operation of layer 4 decoder 522, i_g and i_E are received and demultiplexed. Gain codeword i_g and the layer 3 error vector $\hat{\mathbf{E}}_3$ are used as input to
 30 the frequency selective gain generator 616 to produce gain vector $\mathbf{g}^*$ according to the corresponding method of encoder 512. Gain vector $\mathbf{g}^*$ is then applied to the layer 3 reconstructed audio vector $\hat{\mathbf{S}}_3$ within scaling unit 618, the output of which is then combined with the layer 4 enhancement layer error vector $\mathbf{E}^*$, which was obtained from error signal decoder 612 through decoding of codeword i_E , to produce the layer
 35 4 reconstructed audio output $\hat{\mathbf{S}}_4$.

FIG. 7 is a flow chart showing the operation of an encoder according to the first and second embodiments of the present invention. As discussed above, both embodiments utilize an enhancement layer that scales the encoded audio with a plurality of scaling values and then chooses the scaling value resulting in a lowest error. However, in the second embodiment of the present invention, frequency selective gain generator 603 is utilized to generate the gain values.

The logic flow begins at step 701 where a core layer encoder receives an input signal to be coded and codes the input signal to produce a coded audio signal. Enhancement layer encoder 406 receives the coded audio signal ($s_c(n)$) and scaling unit 401 scales the coded audio signal with a plurality of gain values to produce a plurality of scaled coded audio signals, each having an associated gain value. (step 703). At step 705, error signal generator 402 determines a plurality of error values existing between the input signal and each of the plurality of scaled coded audio signals. Gain selector 404 then chooses a gain value from the plurality of gain values (step 707). As discussed above, the gain value ($\mathbf{g}^*$) is associated with a scaled coded audio signal resulting in a low error value ($\mathbf{E}^*$) existing between the input signal and the scaled coded audio signal. Finally at step 709 transmitter 418 transmits the low error value ($\mathbf{E}^*$) along with the gain value ($\mathbf{g}^*$) as part of an enhancement layer to the coded audio signal. As one of ordinary skill in the art will recognize, both $\mathbf{E}^*$ and $\mathbf{g}^*$ are properly encoded prior to transmission.

As discussed above, at the receiver side, the coded audio signal will be received along with the enhancement layer. The enhancement layer is an enhancement to the coded audio signal that comprises the gain value ($\mathbf{g}^*$) and the error signal ($\mathbf{E}^*$) associated with the gain value.

While the invention has been particularly shown and described with reference to a particular embodiment, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the invention. For example, while the above techniques are described in terms of transmitting and receiving over a channel in a telecommunications system, the techniques may apply equally to a system which uses the signal compression system for the purposes of reducing storage requirements on a digital media device, such as a solid-state memory device or computer hard disk. It is intended that such changes come within the scope of the following claims.

Claims

1. A method for embedded coding of a signal, comprising the steps of:
 receiving an input signal to be coded;
 5 coding the input signal to produce a coded audio signal;
 scaling the coded audio signal with a plurality of gain values to produce a plurality of scaled coded audio signals, each having an associated gain value;
 determining a plurality of error values based on the input signal and each of the plurality of scaled coded audio signals;
 10 choosing a gain value from the plurality of gain values; and
 transmitting or storing the gain value as part of an enhancement layer to the coded audio signal.
2. The method of claim 1 wherein the plurality of gain values comprise frequency
 15 selective gain values.
3. The method of claim 1 wherein the plurality of gain values are a function of a previously encoded signal layer.
- 20 4. A method for receiving a coded audio signal and an enhancement to the coded audio signal, the method comprising the steps of:
 receiving the coded audio signal; and
 receiving the enhancement to the coded audio signal, wherein the enhancement to the coded audio signal comprises a gain value and an error signal
 25 associated with the gain value, wherein the gain value was chosen by a transmitter from a plurality of gain values, wherein the gain value is associated with a scaled coded audio signal resulting in a particular error value existing between an audio signal and the scaled coded audio signal; and
 enhancing the coded audio signal based on the gain value and the error value.
 30
5. The method of claim 4 wherein the gain value comprises a frequency selective gain value.
- 35 6. The method of claim 5 wherein the frequency selective gain values

$$g_j(k) = \begin{cases} \alpha; & k_s \leq k \leq k_e \\ \gamma_j(k); & \text{otherwise} \end{cases},$$

where generally $0 \leq \gamma_j(k) \leq 1$ and $g_j(k)$ is the gain of a k -th position of a j -th candidate vector.

7. An apparatus comprising:

- 5 an encoder receiving an input signal to be coded and coding the input signal to produce a coded audio signal;
- a scaling unit scaling the coded audio signal with a plurality of gain values to produce a plurality of scaled coded audio signals, each having an associated gain value;
- 10 an error signal generator determining plurality of error values existing between the input signal and each of the plurality of scaled coded audio signals;
- a gain selector choosing a gain value from the plurality of gain values, wherein the gain value is chosen based on the error values existing between the input signal and the scaled coded audio signal; and
- 15 a transmitter transmitting the low error value along with the gain value as part of an enhancement layer to the coded audio signal.

8. An apparatus comprising:

- a decoder receiving a coded audio signal; and
- 20 an enhancement layer decoder receiving enhancement to the coded audio signal and producing an enhanced audio signal, wherein the enhancement to the coded audio signal comprises a gain value and an error signal associated with the gain value, wherein the gain value was chosen by a transmitter from a plurality of gain values, wherein the gain value is associated with a scaled coded audio signal resulting in a
- 25 particular error value existing between an audio signal and the scaled coded audio signal.

9. An apparatus comprising:

- a decoder receiving codewords to produce a coded audio signal; and
- 30 an enhancement layer decoder receiving codewords for enhancement to the coded audio signal and outputting an enhanced coded audio signal, wherein the enhancement to the coded audio signal comprises a frequency selective gain value and an error signal associated with the gain value, wherein the frequency selective gain value is based on the coded audio signal.

35

10. A method for decoding a multi-layer encoded audio signal, the method comprising the steps of:

- receiving a first reconstructed audio vector $\hat{\mathbf{S}}_3$ from a first signal decoder;
- receiving a first frequency domain error vector $\hat{\mathbf{E}}_3$ from a first enhancement layer decoder;
- generating a frequency selective gain vector $\mathbf{g}^*$ based on at least the first
- 5 frequency domain error vector;
- scaling the first reconstructed audio signal with the frequency selective gain vector to produce a scaled reconstructed audio signal;
- receiving a codeword i_E for input to a second enhancement layer decoder to produce a second enhancement layer error vector $\mathbf{E}^*$; and
- 10 combining the scaled reconstructed audio signal with the second enhancement layer error vector to produce a decoded multi-layer audio signal output $\hat{\mathbf{S}}_4$.

AMENDED CLAIMS**received by the International Bureau on 20 January 2009 (20.01.2009)**

1. A method for an audio encoder to embed coding of a signal, comprising the steps of:

the audio encoder receiving an input signal to be coded;

the audio encoder coding the input signal to produce a reconstructed audio signal;

the audio encoder scaling the reconstructed audio signal with a plurality of gain values to produce a plurality of scaled reconstructed audio signals, each having an associated gain value;

the audio encoder determining a plurality of error values based on the input signal and each of the plurality of scaled reconstructed audio signals;

the audio encoder choosing a gain value from the plurality of gain values based on the plurality of error values; and

the audio encoder transmitting or storing the gain value as part of an enhancement layer to a coded audio signal.

2. The method of claim 1 wherein the plurality of gain values comprise frequency selective gain values.

3. The method of claim 1 wherein the plurality of gain values are a function of a previously encoded signal layer.

4. A method for an audio decoder receiving a coded audio signal and an enhancement to the coded audio signal, the method comprising the steps of:

the audio decoder receiving the coded audio signal; and

the audio decoder receiving the enhancement to the coded audio signal, wherein the enhancement to the coded audio signal comprises a gain value and an error signal associated with the gain value, wherein the gain value was chosen by a transmitter from a plurality of gain values, wherein the gain value is associated with a scaled reconstructed audio signal resulting in a particular error value existing between an audio signal and the scaled reconstructed audio signal; and

the audio decoder enhancing the coded audio signal based on the gain value and the error value.

5. The method of claim 4 wherein the gain value comprises a frequency selective gain value.

6. The method of claim 5 wherein the frequency selective gain values

$$g_j(k) = \begin{cases} \alpha; & k_s \leq k \leq k_e \\ \gamma_j(k); & \text{otherwise} \end{cases}$$

where generally $0 \leq \gamma_j(k) \leq 1$ and $g_j(k)$ is the gain of a k -th position of a j -th candidate vector.

7. An apparatus comprising:

an encoder receiving an input signal to be coded and coding the input signal to produce a reconstructed audio signal;

a scaling unit scaling the reconstructed audio signal with a plurality of gain values to produce a plurality of scaled reconstructed audio signals, each having an associated gain value;

an error signal generator determining plurality of error values existing between the input signal and each of the plurality of scaled reconstructed audio signals;

a gain selector choosing a gain value from the plurality of gain values, wherein the gain value is chosen based on the plurality of error values existing between the input signal and the scaled reconstructed audio signal; and

a transmitter transmitting the selected gain value as part of an enhancement layer to a coded audio signal.

8. The apparatus of claim 7 wherein the plurality of gain values comprise frequency selective gain values.

9. The apparatus of claim 8 wherein the frequency selective gain values

$$g_j(k) = \begin{cases} \alpha; & k_s \leq k \leq k_e \\ \gamma_j(k); & \text{otherwise} \end{cases}$$

where generally $0 \leq \gamma_j(k) \leq 1$ and $g_j(k)$ is the gain of a k -th position of a j -th candidate vector.

10. An apparatus comprising:

a decoder receiving a coded audio signal; and

an enhancement layer decoder receiving enhancement to the coded audio signal and producing an enhanced audio signal, wherein the enhancement to the coded audio signal comprises a gain value and an error signal associated with the gain value, wherein the gain value was chosen by an encoder from a plurality of gain values, wherein the gain value is associated with a scaled reconstructed audio signal resulting in a particular error value existing between an input audio signal and the scaled reconstructed audio signal.

11. An apparatus comprising:

a decoder receiving codewords to produce a reconstructed audio signal; and

an enhancement layer decoder receiving codewords for enhancement to the coded audio signal and outputting an enhanced reconstructed audio signal, wherein the enhancement to the reconstructed audio signal comprises a frequency selective gain value and an error signal associated with the gain value, wherein the frequency selective gain value is based on the reconstructed audio signal.

12. A method for a decoder to decode a multi-layer encoded audio signal, the method comprising the steps of:

the decoder receiving a first reconstructed audio vector $\hat{S}_3$ from a first signal decoder;

the decoder receiving a first frequency domain error vector $\hat{E}_3$ from a first enhancement layer decoder;

the decoder generating a frequency selective gain vector $\mathbf{g}^*$ based on at least the first frequency domain error vector;

the decoder scaling the first reconstructed audio signal with the frequency selective gain vector to produce a scaled reconstructed audio signal;

the decoder receiving a codeword i_E for input to a second enhancement layer decoder to produce a second enhancement layer error vector $\mathbf{E}^*$; and

the decoder combining the scaled reconstructed audio signal with the second enhancement layer error vector to produce a decoded multi-layer audio signal output $\hat{S}_4$.

13. The method of claim 12 wherein the frequency domain comprises the MDCT domain.

14. The method of claim 12 wherein the step of generating the frequency selective gain vector further comprises:

receiving a gain codeword i_g ; and

generating the frequency selective gain vector based on the gain codeword and the first frequency domain error vector.

15. The method of claim 12 wherein the frequency selective gain vector comprises $g_j(k)$, wherein $g_j(k)$ is the gain of a k -th frequency component of a j -th candidate vector.

100
- PRIOR ART -

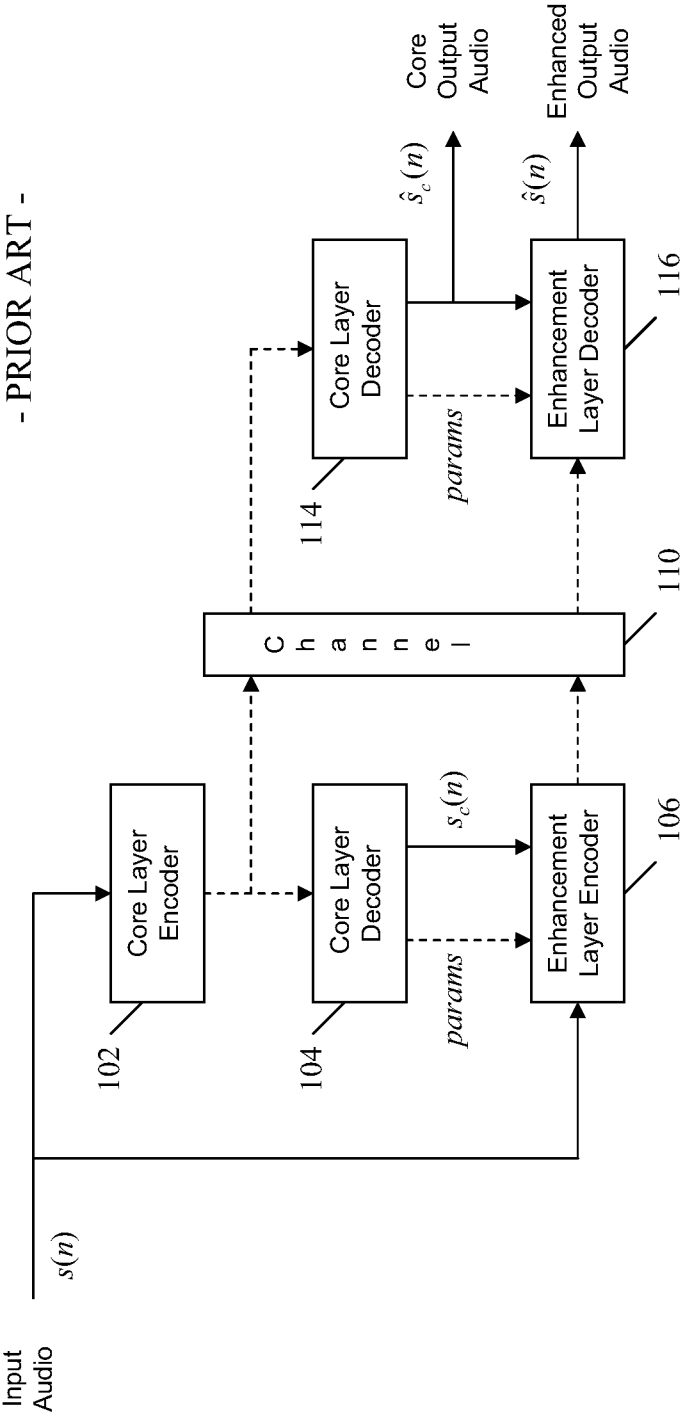


FIG. 1

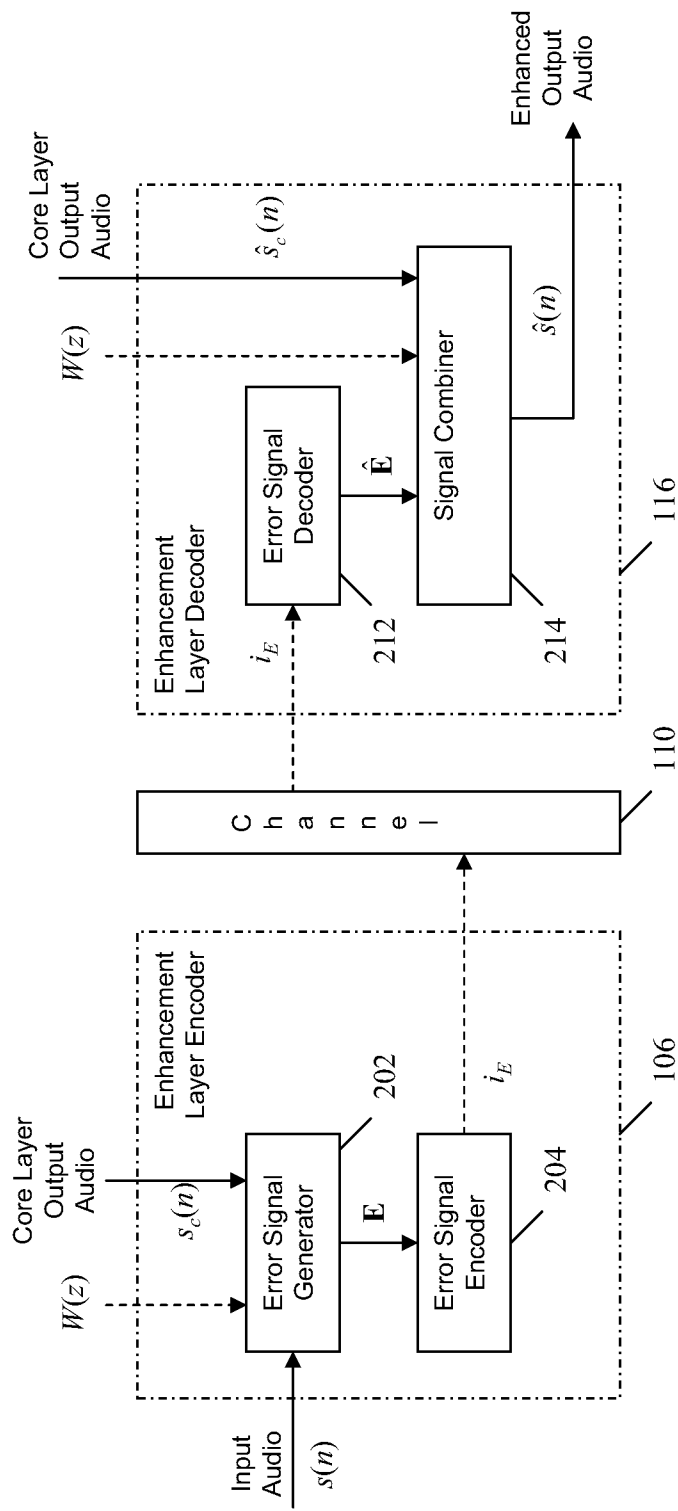


FIG. 2

200

- PRIOR ART -

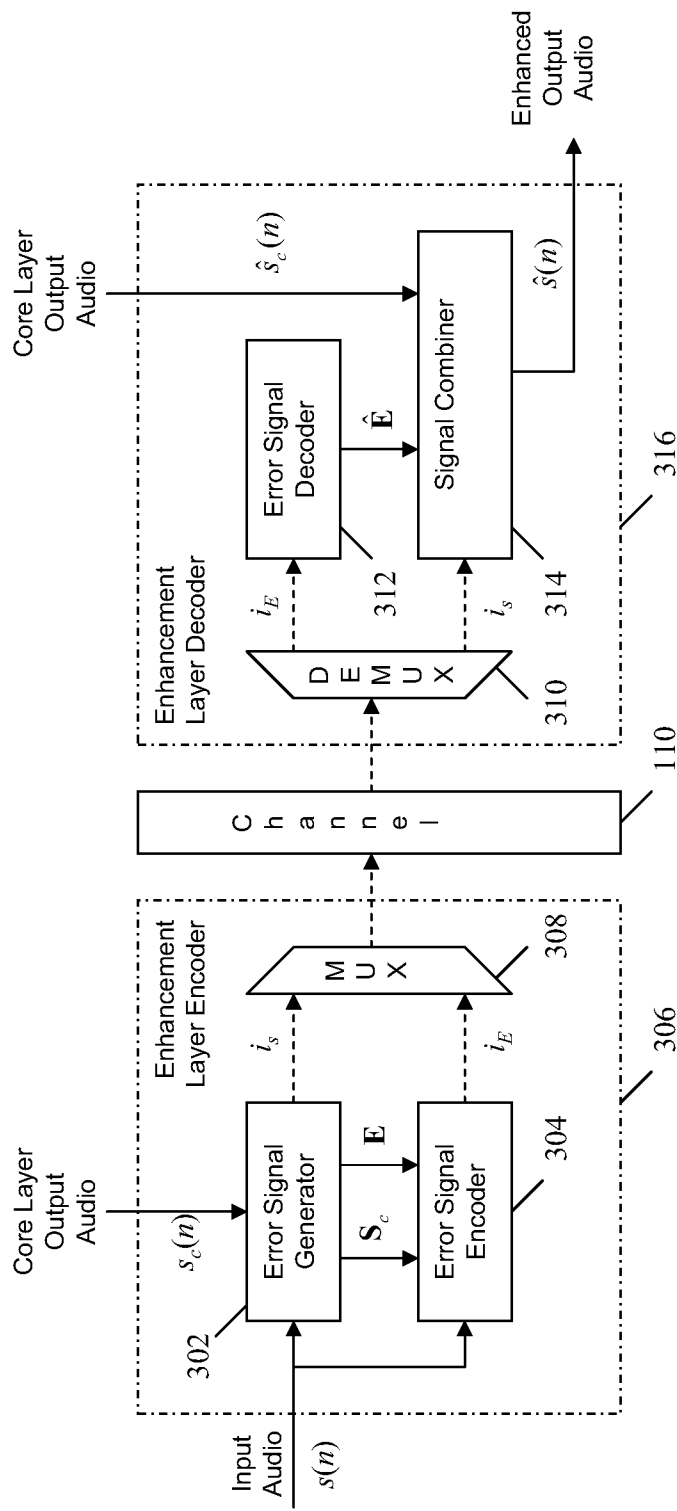


FIG. 3

300

- PRIOR ART -

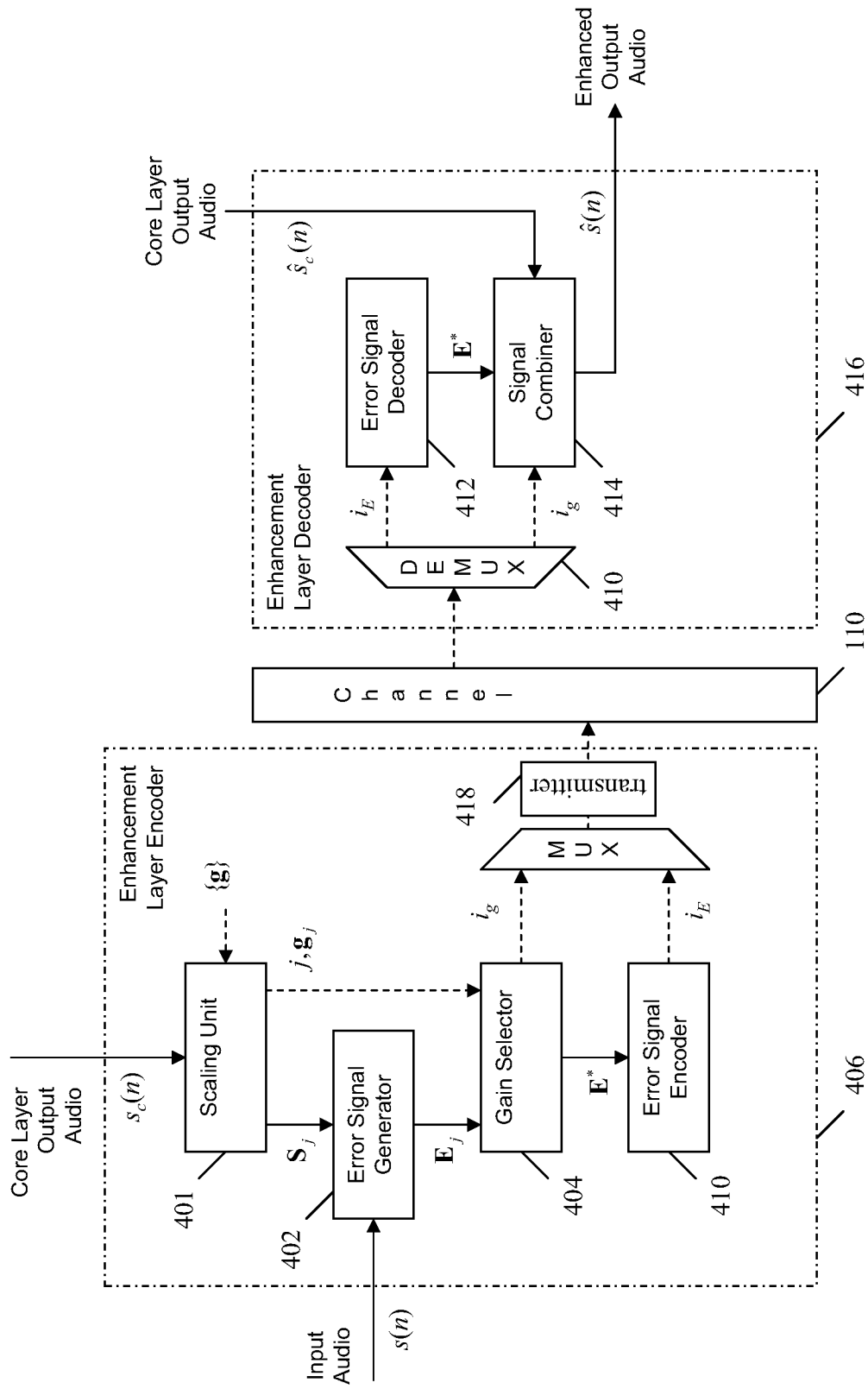


FIG. 4

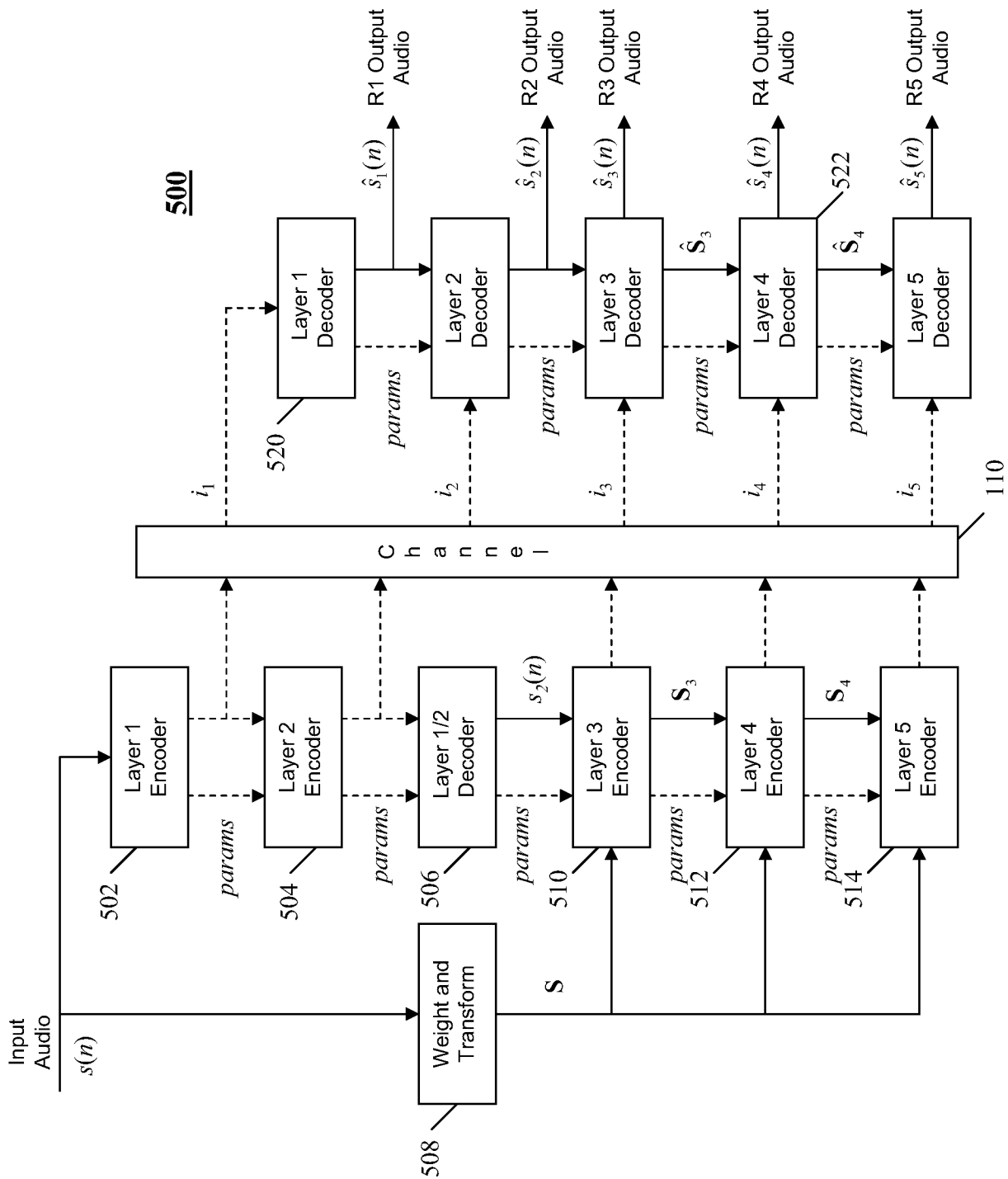


FIG. 5

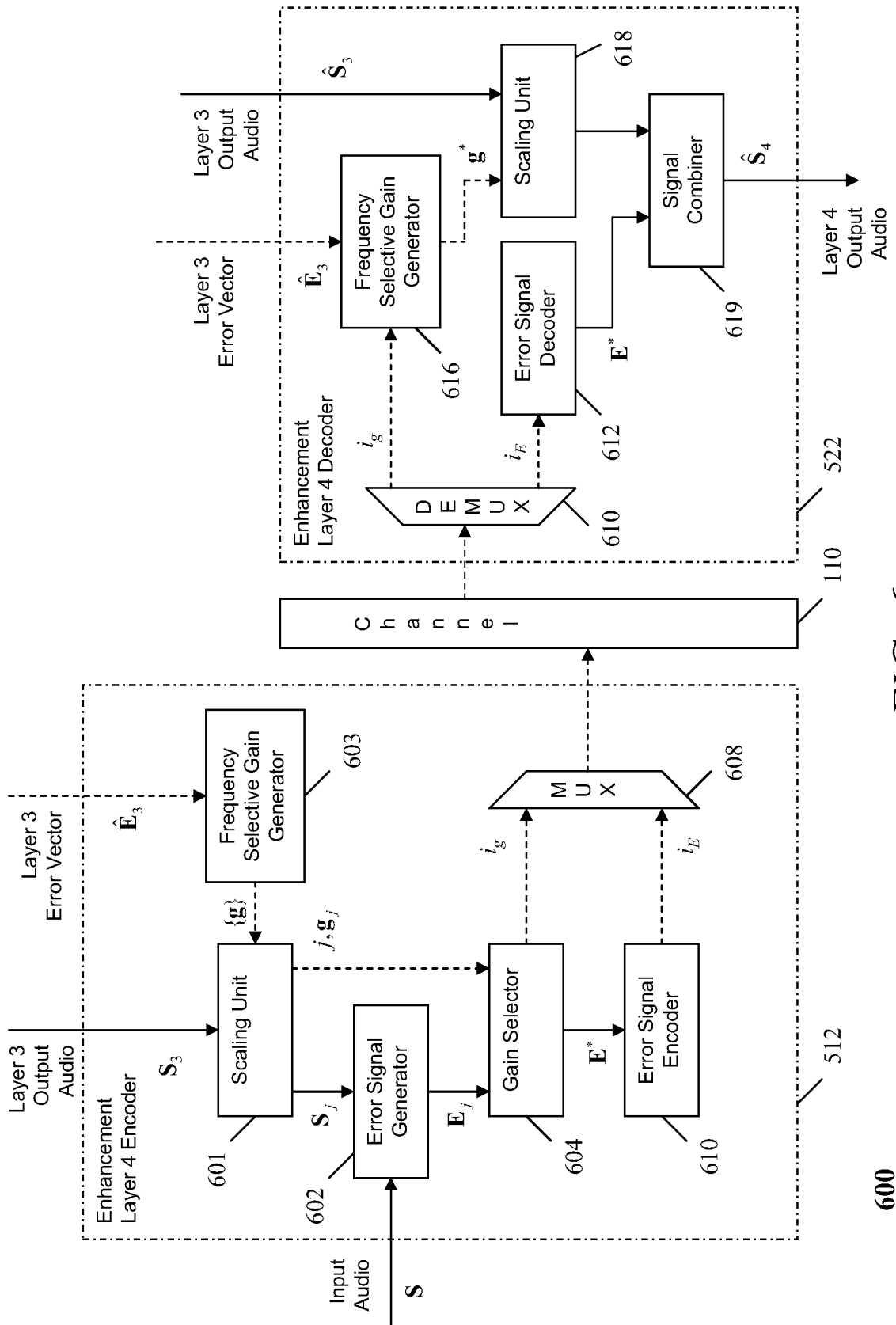


FIG. 6

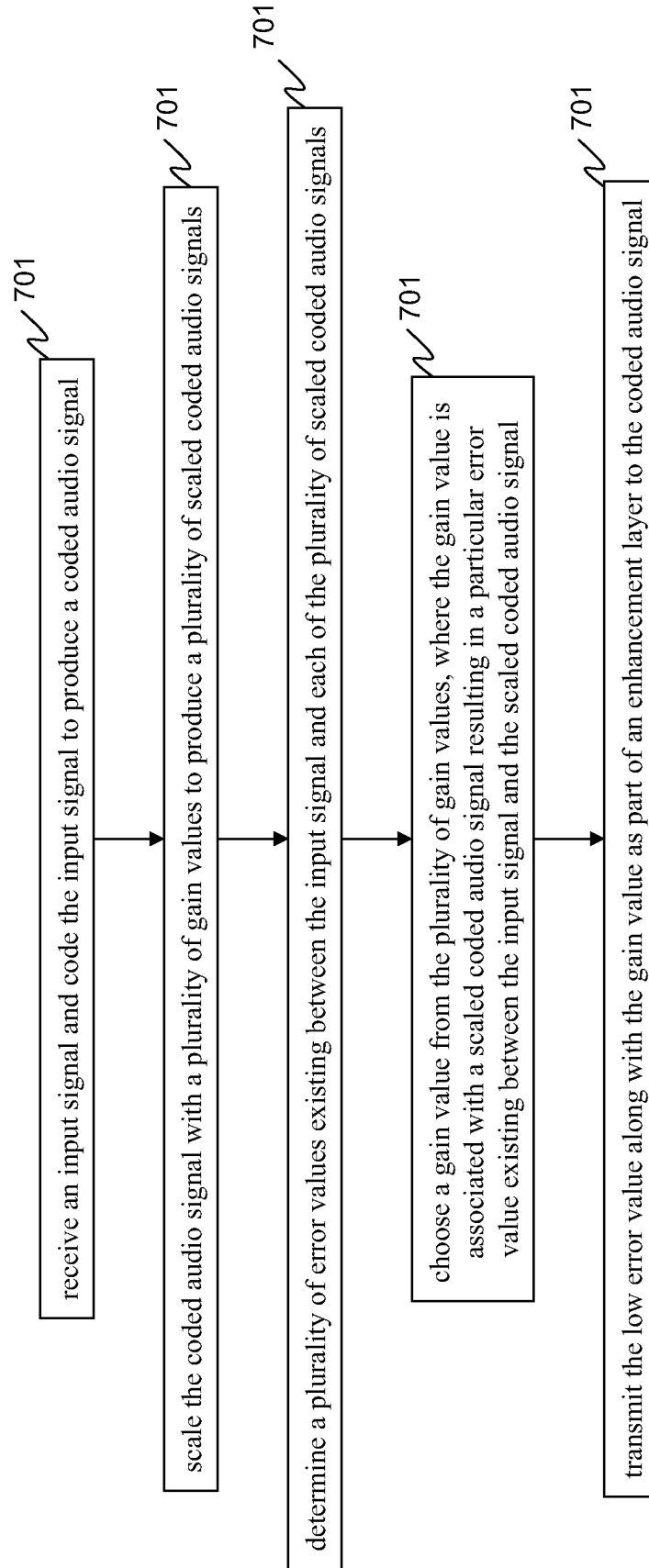


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2008/077693

A. CLASSIFICATION OF SUBJECT MATTER

INV. G10L19/14

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

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

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

EPO-Internal, WPI Data, INSPEC, COMPENDEX, IBM-TDB

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KOVESI B ET AL: "A scalable speech and audio coding scheme with continuous bitrate flexibility" ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2004. PROCEEDINGS. (ICASSP ' 04). IEEE INTERNATIONAL CONFERENCE ON MONTREAL, QUEBEC, CANADA 17-21 MAY 2004, PISCATAWAY, NJ, USA, IEEE, vol. 1, 17 May 2004 (2004-05-17), pages 273-276, XP010717618 ISBN: 978-0-7803-8484-2 abstract paragraph [0005]; figure 4 ----- -/--	1-10

☒ 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
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

- *T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- *&* document member of the same patent family

Date of the actual completion of the international search

14 November 2008

Date of mailing of the international search report

15/12/2008

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
 NL - 2280 HV Rijswijk
 Tel. (+31-70) 340-2040,
 Fax: (+31-70) 340-3016

Authorized officer

Zimmermann, Elko

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2008/077693

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	RAMPRASHAD S A: "Embedded coding using a mixed speech and audio coding paradigm" INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY KLUWER ACADEMIC PUBLISHERS NETHERLANDS, vol. 2, no. 4, May 1999 (1999-05), pages 359-372, XP002503923 ISSN: 1381-2416 abstract; figures 1,6,7 page 365, right-hand column, line 14 - page 366, right-hand column, line 13 page 367, left-hand column, line 14 - right-hand column, line 5 page 368, right-hand column, line 32 - page 369, left-hand column, line 21 -----	1-10