

Relatório Descritivo da Patente de Invenção para **"APERFEIÇOAMENTO DA FALA EM AUDIO DE ENTRETENIMENTO"**.

Descrição

Campo técnico

5 A presente invenção refere-se a processamento de sinal de áudio. Mais especificamente, a invenção refere-se a processamento de áudio de entretenimento, como áudio de televisão, para aperfeiçoar a clareza e inteligibilidade da fala, tais como diálogo e narrativa de áudio. A invenção refere-se a métodos, a aparelho para executar tais métodos, e a software
10 armazenado em um meio legível de computador para causar a um computador a executar tais métodos.

Antecedentes da Técnica

 O entretenimento audiovisual evoluiu dentro de uma rápida sequência compassada de diálogo, narrativa, música, e efeitos. O alto realismo
15 realizável com tecnologias modernas de áudios de entretenimento e métodos de saída encorajou o uso de estilos conversacionais de falar na televisão que diferem substancialmente da apresentação como estágio anunciado claramente do passado. Esta situação coloca um problema não somente para a população crescente de espectadores de idade avançada que, em
20 face da diminuição sensorial e das habilidades de processamento de linguagem, devem se esforçar para seguir a programação mas também para pessoas com audição normal, por exemplo, quando escutando em níveis acústicos baixos.

 Como a fala é entendida depende de vários fatores. Exemplos
25 são o cuidado de saída da fala (fala clara ou conversacional), a taxa da fala, e a audibilidade da fala. O idioma falado é notavelmente robusto e pode ser entendido sob condições menores do que as ideais. Por exemplo, ouvintes de audição prejudicada podem, tipicamente, seguir fala clara até mesmo quando eles não puderem ouvir partes da fala devido a acuidade de audição
30 diminuída. No entanto, enquanto a taxa da fala aumenta e a saída da fala torna-se menos precisa, escutar e compreender exige esforço crescente, particularmente se partes do espectro da fala são inaudíveis.

Devido ao fato de que públicos de televisão não podem fazer nada para afetar a clareza da fala de radiodifusão, os ouvintes de audição prejudicada podem tentar compensar a audibilidade inadequada aumentando o volume de escutar. Aparte de ser censurável para pessoas de audição normal na mesma sala ou para vizinhos, esta abordagem é somente parcialmente eficaz. Isto é assim porque a maior parte das perdas de audição não são uniformes através das frequências; elas afetam frequências altas mais do que baixas e médias frequências. Por exemplo, a habilidade típica de um macho de 70 anos de ouvir sons a 6 kHz é mais ou menos 50 dB pior do que aquela de uma pessoa jovem, mas em frequências abaixo de 1 kHz a desvantagem de audição da pessoa mais velha é menos do que 10 dB (ISO 7029, Acoustics – Statistical distribution of hearing thresholds as a function of age). Aumentar o volume torna sons de frequência baixa e média mais altos sem aumentar significativamente sua contribuição para a inteligibilidade porque para aquelas frequências a audibilidade já é adequada. Aumentar o volume também faz pouco para superar a perda de audição significativa em frequências altas. Uma correção mais apropriada é um controle de tom, como que provido por um equalizador gráfico.

Embora uma melhor opção do que simplesmente aumentar o controle de volume, um controle de tom é ainda insuficiente para a maior parte das perdas de audição. O grande ganho requerido de frequência alta para tornar passagens suaves audíveis para o ouvinte de audição prejudicada é provável ser desconfortavelmente alto durante as passagens de nível alto e pode até sobrecarregar a cadeia de reprodução de audio. Uma solução melhor é amplificar dependendo do nível do sinal, provendo ganhos maiores para partes de sinal de baixo nível e ganhos menores (ou nenhum ganho mesmo) para partes de nível alto. Tais sistemas, conhecidos como controles de ganho automático (AGC) ou compressores dinâmicos de faixa (DRC) são usados na ajuda de audição e seu uso para aperfeiçoar a inteligibilidade para a audição prejudicada em sistemas de telecomunicação foi proposta (por exemplo, patente US 5,388,185, Patente US 5,539,806, e Patente US 6,061,431).

Devido a que perda de audição desenvolve-se geralmente gradualmente, a maioria dos ouvintes com dificuldades de audição cresceu acostumado a suas perdas. Como resultado, eles frequentemente objetam a qualidade de som de audio de entretenimento quando ele é processado para compensar por sua deterioração de audição. Públicos de audição prejudicada são mais prováveis de aceitar a qualidade de som de audio compensado quando ele provê um benefício tangível para eles, tal como quando ele aumenta a inteligibilidade de diálogo e narrativa ou reduz o esforço mental requerido para compreensão. Portanto é vantajoso limitar a aplicação compensação de perda de audição para aquelas partes do programa de audio que são dominados por fala. Fazendo assim, otimiza a escolha entre modificações de qualidade de som potencialmente censuráveis de música e sons ambientes por um lado e os benefícios de inteligibilidade desejáveis no outro.

15 Descrição da Invenção

De acordo com um aspecto da invenção, a fala no audio de entretenimento pode ser aperfeiçoada processando, em resposta a um ou mais controles, o audio de entretenimento para aperfeiçoar a clareza e inteligibilidade de partes da fala do audio de entretenimento, gerar um controle para o processamento, a geração incluindo caracterizar segmentos de tempo do audio de entretenimento como (a) fala ou não-fala ou (b) como provável de ser fala ou não-fala, e responder a mudanças no nível do audio de entretenimento para prover um controle para o processamento, em que tais mudanças são respondidas dentro de um período de tempo menor do que os segmentos de tempo, e um critério de decisão da resposta é controlado pela caracterização. Cada um do processamento e da resposta podem operar em faixas de frequência múltipla correspondente, a resposta provendo um controle para o processamento para cada uma das faixas de frequência múltipla.

Aspectos da invenção podem operar em uma maneira de "olhar adiante" tal que quando existe acesso a uma evolução de tempo do audio de entretenimento antes e depois de um ponto de processamento, e em que o gerar um controle responde a pelo menos algum audio depois do ponto de

processamento.

Aspectos da invenção podem empregar separação temporal e/ou espacial tais que passos do processamento, como caracterizar e responder são executados em tempos diferentes ou em lugares diferentes. Por exemplo, a caracterização pode ser executada em um primeiro tempo ou lugar, o processamento e resposta podem ser executados em um segundo tempo ou lugar, e informação sobre a caracterização de segmentos de tempo pode ser armazenada ou transmitida para controlar os critérios de decisão da resposta.

Aspectos da invenção podem incluir codificar também o audio de entretenimento conforme um esquema de codificação perceptiva ou um esquema de codificação sem perda, e decodificar o audio de entretenimento de acordo com o mesmo esquema de codificação empregado pela codificação, em que passos do processamento, como caracterizar e responder são executados junto com a codificação ou a decodificação. A caracterização pode ser executada junto com a codificação e o processamento e/ou o resposta pode ser executado junto com a decodificação.

De acordo com aspectos acima mencionados da invenção, o processamento pode operar de acordo com um ou mais parâmetros de processamento. O ajuste de um ou mais parâmetros podem ser em resposta ao audio de entretenimento de tal forma que uma métrica de inteligibilidade da fala do audio processado é ou maximizado ou impulsionado acima de um nível de limite desejado. De acordo com aspectos da invenção, o audio de entretenimento pode compreender canais múltiplos de audio nos quais um canal é principalmente fala e o um ou mais outros canais são principalmente de não-fala, em que a métrica da inteligibilidade de fala é baseada no nível do canal da fala e o nível no um ou mais outros canais. A métrica da inteligibilidade de fala pode ser também baseada no nível de ruído no ambiente de escutar em que o audio processado é reproduzido. O ajuste de um ou mais parâmetros pode ser em resposta a um ou mais descritos a longo prazo do audio de entretenimento. Os exemplos de descritores a longo prazo incluem o nível médio de diálogo do audio de entretenimento e uma estimativa de

processamento já aplicado ao audio de entretenimento. O ajuste de um ou mais parâmetros podem ser de acordo com uma fórmula prescritiva, em que a fórmula prescritiva relaciona a acuidade de audição de um ouvinte ou grupo de ouvintes a um ou mais parâmetros. Alternativamente, ou adicionalmente, o ajuste de um ou mais parâmetros pode ser de acordo com as preferências de um ou mais ouvintes.

De acordo com os aspectos acima mencionados da invenção o processamento pode incluir funções múltiplas atuando em paralelo. Cada uma das funções múltiplas pode operar em uma das faixas de frequência múltipla. Cada uma das funções múltiplas pode fornecer, individualmente ou coletivamente, controle dinâmico de faixa, equalização dinâmica, estreitamento espectral, transposição de frequência, extração de fala, redução de ruído, ou outra ação de aperfeiçoamento de fala. Por exemplo, o controle dinâmico de faixa pode ser provido por funções ou dispositivos múltiplos de compressão/expansão, em que cada um processa uma região de frequência do sinal de audio.

A parte do processamento inclui ou não funções múltiplas atuando em paralelo, o processo pode prover controle dinâmico de faixa, equalização dinâmica, estreitamento espectral, transposição de frequência, extração de fala, redução de ruído, ou outra ação de aperfeiçoamento de fala. Por exemplo, o controle dinâmico de faixa pode ser provido por uma função ou dispositivo de compressão/expansão dinâmico de faixa.

Um aspecto da invenção é controlar o aperfeiçoamento da fala adequado para compensação de perda de audição de tal forma que, idealmente, opera somente nas partes da fala de um programa de audio e não opera nas partes restante (não-fala) de programa, assim tendendo a não mudar o timbre (distribuição espectral) ou sonoridade percebida das partes restantes (não-fala) do programa.

De acordo com outro aspecto da invenção, aperfeiçoar a fala em audio de entretenimento inclui analisar o audio de entretenimento para classificar segmentos de tempo do audio como sendo fala ou outro audio, e aplicar compressão dinâmica de faixa a uma ou múltiplas faixas de frequência

do audio de entretenimento durante segmentos de tempo classificados como fala.

Descrição Dos Desenhos

5 A Figura 1a é um diagrama de bloco funcional esquemático ilustrando uma implementação exemplificativa de aspectos da invenção.

A Figura 1b é um diagrama de bloco funcional esquemático mostrando uma implementação exemplificativa de uma versão modificada da Figura 1a na qual dispositivos e/ou funções podem ser separados temporalmente e/ou espacialmente.

10 A Figura 2 é um diagrama de bloco funcional esquemático mostrando uma implementação exemplificativa de uma versão modificada da Figura 1a na qual o controle de aperfeiçoamento da fala é derivado de uma maneira de "olhar adiante".

15 As Figuras 3a-c são exemplos de transformações de ganho de potência úteis no entendimento do exemplo da Figura 4.

A Figura 4 é um diagrama de bloco funcional esquemático mostrando como o ganho no aperfeiçoamento da fala em uma faixa de frequência pode ser derivado da estimativa de potência do sinal daquela faixa de acordo com aspectos da invenção.

20 Melhor Modo Para Executar a Invenção.

As técnicas para classificar audio em fala e não-fala (como música) são conhecidas na técnica e são às vezes conhecidas como um discriminador de fala-versus-outro ("SVO"). Ver, por exemplo, as Patentes US 6,785,645 e 6,570,991 como também os Pedidos de Patente Publicados US 20040044525, e as referências contidas nelas. Os discriminadores de audio 25 fala-versus-outros analisam segmentos de tempo de um sinal de audio e extraem um ou mais descritores de sinal (características) de todo segmento de tempo. Tais características são passadas a um processador que tanto produz uma estimativa de probabilidade do segmento de tempo ser fala, 30 quanto toma uma árdua decisão de fala/não-fala. A maioria das características refletem a evolução de um sinal com o passar do tempo. Exemplos típicos de características são a taxa na qual o espectro de sinal muda com o

passar do tempo ou a inclinação da distribuição da taxa na qual a polaridade do sinal muda. Para refletir as características distintas da fala confiavelmente, os segmentos de tempo devem ser de comprimento suficiente. Devido a que muitas características são baseadas em características de sinal que refletem as transições entre sílabas adjacentes, os segmentos de tempo tipicamente cobrem pelo menos a duração de duas sílabas (isto é, mais ou menos 250 ms) para capturar uma tal transição. Porém, segmentos de tempo são frequentemente mais longos (por exemplo, por um fator de cerca de 10) obter estimativas mais confiáveis. Embora relativamente lento em operação, os SVOs são razoavelmente confiáveis e precisos em classificar audio em fala e não-fala. Porém, para aperfeiçoar a fala seletivamente, em um programa de audio de acordo com aspectos da presente invenção, é desejável controlar o aperfeiçoamento da fala em uma escala de tempo mais fina do que a duração dos segmentos de tempo analisados por um discriminador de fala-versus-outro.

Outra classe de técnicas, algumas vezes conhecidas como detectores de atividade de voz (VADs) indicam a presença ou ausência da fala em um fundo de ruído relativamente estável. Os VADs são usados extensivamente como parte de esquemas de redução de ruído em aplicações de comunicação da fala. Diferentemente dos discriminadores fala-versus-outros, os VADs normalmente tem uma resolução temporal que é adequada para o controle de aperfeiçoamento da fala conforme aspectos da presente invenção. Os VADs interpretam um aumento súbito da potência de sinal como o início de um som da fala e uma diminuição súbita da potência de sinal como o fim de um som de fala. Fazendo assim, eles sinalizam a demarcação entre fala e fundo quase instantaneamente (isto é, dentro de uma janela de integração temporal para medir a potência de sinal, por exemplo, mais ou menos 10 ms). No entanto, devido a que os VADs reagem a qualquer mudança súbita da potência de sinal, eles não podem diferenciar entre fala e outros sinais dominantes, tal como música. Portanto, se usados sózinhos, os VADs não são adequados para controlar o aperfeiçoamento da fala para aperfeiçoar a fala seletivamente de acordo com a presente invenção.

É um aspecto da invenção combinar a especificidade fala versus não-fala de discriminadores fala-versus-outro (SVO) com a acuidade temporal de detectores de atividade de voz (VADs) para facilitar o aperfeiçoamento da fala que responde seletivamente a fala em um sinal de audio com uma resolução temporal que é mais fina do que a encontrada em discriminadores de fala-versus-outro da técnica anterior.

Embora, em princípio, aspectos da invenção possam ser implementados em domínios analógicos e/ou digitais, implementações práticas são prováveis de serem implementadas no domínio digital em que cada um dos sinais de audio são representados por amostras individuais ou amostras dentro de blocos de dados.

Referindo-se agora à Figura 1a, é mostrado um diagrama de bloco funcional esquemático ilustrando aspectos da invenção na qual um sinal de entrada de audio 1 é passado para uma função ou dispositivo de aperfeiçoamento da fala ("Aperfeiçoamento da Fala ") 102 que, quando habilitado por um sinal de controle 103, produz um sinal de saída de audio de fala aperfeiçoada 104. O sinal de controle é gerado por uma função ou dispositivo de controle "Controlador de Aperfeiçoamento da Fala") 105 que opera em segmentos de tempo armazenados em buffer do sinal de entrada de áudio 101. O Controlador de Aperfeiçoamento da Fala 105 inclui uma função ou dispositivo discriminador de fala-versus-outro ("SVO") 107 e um conjunto de um ou mais funções ou dispositivos detectores de atividade ("VAD") 108. O SVO 107 analisa o sinal ao longo de uma duração de tempo que é mais longa do que aquela analisada pelo VAD. O fato de que o SVO 107 e o VAD 108 operam ao longo do tempo com durações de tempo de comprimentos diferentes é ilustrado por meio da ilustração de um parêntese acessando uma região larga (associado com o SVO 107) e outro parêntese acessando uma região mais estreita (associado com o VAD 108) de uma função ou dispositivo de armazenamento de ("Buffer") 106. A região larga e a região mais estreita são esquemáticas e não em escala. No caso de uma implementação digital em que os dados de audio são transportados em blocos, cada parte do Buffer 106 pode armazenar um bloco de dados de audio. A região aces-

sada pelo VAD inclui as partes mais recentes do armazenamento de sinal no Buffer 106. A probabilidade da seção de sinal atual ser fala, como determinado pelo SVO 107, serve para controlar 109 o VAD 108. Por exemplo, ele pode controlar um critério de decisão do VAD 108, assim desviando as decisões do VAD.

O buffer 106 simboliza memória inerente ao processamento e pode ou não pode ser implementada diretamente. Por exemplo, se o processamento é executado em um sinal de audio que é armazenado em um meio com acesso de memória randomica, aquele meio pode servir como buffer. Similarmente, a história da entrada de audio pode ser refletida no estado interno do discriminador de fala-versus-outro 107 e o estado interno do detector de atividade de voz, em cujo caso nenhum buffer separado é necessário.

O aperfeiçoamento da fala 102 pode ser composto de múltiplos dispositivos ou funções de processamento de audio que trabalham em paralelo para aperfeiçoar a fala. Cada dispositivo ou função pode operar em uma região de frequência do sinal de audio em que a fala é para ser aperfeiçoada. Por exemplo, os dispositivos ou funções podem prover, individualmente ou como um todo, controle dinâmico de faixa, equalização dinâmica, estreitamento espectral, transposição de frequência, extração da fala, redução de ruído, ou outra ação de aperfeiçoar fala. Nos exemplos detalhados de aspectos da invenção, o controle dinâmico de faixa provê compressão e/ou expansão em faixas de frequência do sinal de audio. Deste modo, por exemplo, o aperfeiçoamento da Fala 102 pode ser um banco de compressores/expansores de faixa ou funções de compressão/expansão dinâmica, em que cada um processa uma região de frequência do sinal de audio (um compressor/expansor ou função de compressão/expansão de faixa múltipla). A especificidade de frequência disposta por compressão/expansão de faixa multipla é útil não somente porque ela permite costurar o padrão de aperfeiçoamento da fala ao padrão de uma dada perda de audição, mas também porque permite responder ao fato de que em qualquer dado momento pode estar presente fala em uma região de frequência mas ausente em outra.

Para tomar vantagem plena da especificidade de frequência oferecida por compressão de faixa múltipla, cada faixa de compressão/expansão pode ser controlada por seu próprio detector de atividade de voz ou função de detecção de voz. Em tal caso, cada detector de atividade de voz ou função de detecção de voz pode sinalizar atividade de voz na região de frequência associada com a faixa de compressão/expansão que ele controla. Embora existam vantagens no Aperfeiçoamento da Fala 102 sendo composto de vários dispositivos ou funções de processamento de áudio que trabalham em paralelo, simples versões de aspectos da invenção podem empregar um Aperfeiçoamento da Fala 2 que é composto de somente um dispositivo ou função de processamento de áudio.

Até mesmo quando existem muitos detectores de atividade de voz, pode haver somente um discriminador de fala-versus-outro 107 gerando uma saída única 9 para controlar todos os detectores de atividade de voz que estão presentes. A escolha para usar somente um discriminador de fala-versus-outro reflete duas observações. Uma é que a taxa em que o padrão de banda passante da atividade de voz muda com o tempo é tipicamente muito mais rápida do que a resolução temporal do discriminador de fala-versus-outro. A outra observação é que as características usadas pelo discriminador de fala-versus-outro são tipicamente derivadas de características espectrais que podem ser observadas melhor em um sinal de banda larga. Ambas as observações tornam o uso de discriminadores de fala de faixa específica-versus-outro impraticáveis.

Uma combinação de SVO 107 e VAD 108 como ilustrada no Controlador de Aperfeiçoamento da Fala 105 pode ser também usado para propósitos outros do que para aperfeiçoar fala, por exemplo para estimar a sonoridade da fala em um programa de áudio, ou para medir a taxa de fala.

O esquema de aperfeiçoamento da fala acabado de ser descrito pode ser desdobrado de muitas formas. Por exemplo, o esquema inteiro pode ser implementado dentro de uma televisão ou um conversor set-top box para operar sobre o sinal de áudio recebido de uma radiodifusão de televisão. Alternativamente, pode ser integrado com um codificador perceptivo de

audio (por exemplo, AC-3 ou AAC) ou ele pode ser integrado com um codificador de audio sem perda.

Aperfeiçoamento da Fala de acordo com aspectos da presente invenção podem ser executados em tempos diferentes ou em lugares diferentes. Considere um exemplo no qual o aperfeiçoamento da fala é integrado ou associado com um codificador de audio ou processamento de codificação. Em tal caso, a parte do discriminador de fala-versus outro (SVO) 107 do Controlador de Aperfeiçoamento da Fala 105, o qual é frequentemente computavelmente caro, pode ser integrado ou associado com o codificador de audio ou processamento de codificação. A saída 109 do SVO, por exemplo uma bandeira indicando a presença de fala, pode ser embutida na corrente de audio codificada. Tal informação embutida em um fluxo de audio codificado é frequentemente referido como meta dados. O Aperfeiçoamento da fala 102 e o VAD 108 do Controlador do Aperfeiçoamento da Fala 105 pode ser integrado ou associado com um decodificador de audio e operar no audio previamente codificado. O conjunto de um ou mais detectores de atividade de voz (VAD) 108 usa também a saída 109 do discriminador de fala-versus-outro (SVO) 107, que ele extrai do fluxo de audio codificado.

A Figura 1b mostra uma implementação exemplificativa de uma tal versão modificada da Figura 1a. Dispositivos ou funções na Figura 1b que correspondem àquelas na Figura 1 recebem os mesmos números de referência. O sinal de entrada de audio 101 é passado para um codificador ou função de codificação ("codificador") 110 e a um Buffer 106 que cobre a duração de tempo requerida pelo SVO 107. O codificador 110 pode ser parte de um sistema perseptivo ou de codificação sem perda. A saída do codificador 110 é passada para um multiplexador ou função de multiplexar ("Múltiplexador") 112. A saída do SVO (109 na Figura 1a) é mostrada como sendo aplicada 109a ao codificador 110 ou, alternativamente, aplicada 109b ao Multiplexador 112 que recebe também a saída do codificador 110. A saída do SVO, tal como uma bandeira como na Figura 1a, é tanto transportada na saída de fluxo de bits do codificador 110 (como metadados, por exemplo) ou é multiplexado com a saída de codificador 110 para prover um pacote e fluxo

de bits montado 114 para armazenamento ou transmissão a um demultiplexador ou função de demultiplexar ("Demultiplexador") 116 que desempacota o fluxo de bits 114 para passar a um decodificador ou uma função de decodificação 118. Se a saída do SVO 107 foi passada 109b para ao Multiplexador 112, então ela é recebida 109b' do Demultiplexador 116 e passada ao VAD 108. Alternativamente, se a saída do SVO 107 foi passada 109a para o codificador 110, então ela é recebida 109a' do Decodificador 118. Como no exemplo da Figura 1a, o VAD 108 pode compreender múltiplas funções ou dispositivos de atividade de voz. Uma função ou dispositivo de buffer de sinal ("Buffer") 120 alimentado pelo Decodificador 118 que cobre a duração de tempo requerida pelo VAD 108 prove outra alimentação para o VAD 108. A saída do VAD 103 é passada para um Aperfeiçoamento de fala 102 que prove a saída de áudio de fala aperfeiçoada como na Figura 1a. Embora mostrada separadamente por clareza na apresentação, o SVO 107 e/ou Buffer 106 pode ser integrado com o codificador 110. Semelhantemente, embora mostrada separadamente por clareza na apresentação, o VAD 108 e/ou Buffer 120 podem ser integrados com o Decodificador 118 ou Aperfeiçoamento de fala 102.

Se o sinal de áudio a ser processado foi pré-gravado, por exemplo como quando reproduzindo de um DVD em uma casa do consumidor ou quando processando fora de linha em um ambiente de radiodifusão, o discriminador de fala-versus-outra e/ou o detector de atividade de voz podem operar em seções de sinal que incluem partes de sinal que, durante a reprodução, acontece depois da amostra de sinal atual ou bloco de sinal. Isto é ilustrado na Figura 2, onde o buffer de sinal simbólico 201 contém seções de sinal que, durante a reprodução, acontece depois da amostra de sinal atual ou bloco de sinal ("olhar adiante"). Ainda que o sinal não esteve pré-gravado, olhar adiante pode ainda ser usado quando o codificador de áudio tem um atraso de processamento inerente significativo.

Os parâmetros de processamento de Aperfeiçoamento de fala 102 podem ser atualizados em resposta ao sinal de áudio processado em uma taxa que é mais baixa do que a taxa de resposta dinâmica do compres-

sor. Existem vários objetivos que se pode perseguir quando atualizando os parâmetros de processador. Por exemplo, o parâmetro de processamento de função de ganho do processador de aperfeiçoamento de fala pode ser ajustado em resposta ao nível médio da fala do programa para assegurar que a mudança do espectro médio de fala de longo prazo é independente do nível da fala. Para entender o efeito de tal ajuste e a necessidade dele, considere o exemplo seguinte. O aperfeiçoamento da fala é aplicado somente a uma parte de frequência alta de um sinal. Em um dado nível médio da fala, a estimativa de potência 301 da parte de sinal de frequência alta é a média P_1 , onde P_1 é maior do que a potência limite de compressão 304. O ganho associado com esta estimativa de potência é G_1 , que é o ganho médio aplicado à parte de alta frequência do sinal. Devido a que a parte de baixa frequência não recebe nenhum ganho, o espectro médio da fala é conformado para ser G_1 dB mais alto nas frequências altas do que nas frequências baixas. Agora considera-se o que acontece quando o nível médio da fala aumenta por uma certa quantia, ΔL . Um aumento do nível médio da fala por ΔL dB aumenta a estimativa de potência média 301 da parte de sinal de alta frequência para $P_2 = P_1 + \Delta L$. Como pode ser visto da Figura 3a, a estimativa de potência mais alta P_2 dá aumento a um ganho, G_2 que é menor do que G_1 . Consequentemente, o espectro médio da fala do sinal processado mostra a ênfase menor de frequência alta quando o nível médio da entrada é alto do que quando ele é baixo. Devido a que ouvintes compensam para diferenças no nível médio da fala com seu controle de volume, a dependência de nível da ênfase de alta frequência média é indesejável. Pode ser eliminada modificando a curva de ganho das Figuras 3a-c em resposta ao nível médio da fala. As Figuras 3a-c são discutidas abaixo.

Os parâmetros de processamento de Aperfeiçoamento de fala 102 podem ser também ajustados para assegurar que uma métrica de inteligibilidade da fala é ou maximizada ou impulsionada acima de um nível limite desejado. A métrica de inteligibilidade da fala pode ser computada dos níveis relativos do sinal de áudio e um som competindo no ambiente de escutar (tal como ruído de cabina de aeronave). Quando o sinal de áudio é um sinal de

audio de multicanal com fala em um canal e sinais de não-fala nos canais restantes, a métrica de inteligibilidade da fala pode ser computada, por exemplo, dos níveis relativos de todos os canais e a distribuição de energia espectral neles. Métricas de inteligibilidade adequadas são bem-conhecidas [por exemplo, ANSI S3.5-1997 "Method for Calculation of the Speech Intelligibility Index", American National Standards Institute, 1997; ou Musch e Buus, "Using statistical decision theory to predict speech intelligibility" J Model Structure," Journal of the Acoustical Society of America, (2001) 109, pp 2896 - 2909].

Aspectos da invenção mostrados nos diagramas de bloco funcional da Figura 1a e 1b e descritos aqui podem ser implementados como no exemplo das Figuras 3a-c e 4. neste exemplo, a amplificação de compressão de conformar frequência de componentes da fala e liberação do processamento para os componentes de não-fala podem ser realizada através de um processador dinâmico de faixa de multibanda (não mostrado) que implementa ambas as características compressivas e expansivas. Tal processador pode ser caracterizado por um conjunto de funções de ganho. Cada função de ganho relaciona-se à potência de entrada em uma faixa de frequência para um ganho de faixa correspondente, a qual pode ser aplicada aos componentes de sinal naquela faixa. Uma tal relação é ilustrada nas Figuras 3a-c.

Referindo-se à Figura 3a, a estimativa da potência de potência de entrada de faixa 301 é relacionada a um ganho de faixa desejado 302 por uma curva de ganho. Aquela curva de ganho é tomada como o mínimo de duas curvas constituintes. Uma curva constituinte, mostrada pela linha sólida, tem uma característica de compressão com uma relação de compressão apropriadamente escolhida ("CR") 303 para estimativas de potência 301 acima de um limite de compressão 304 e um ganho constante para estimativas de potência abaixo do limite de compressão. A outra curva constituinte, mostrada pela linha tracejada, tem uma característica expansiva com uma relação de expansão apropriadamente escolhida ("ER") 305 para estimativas de potência acima do limite de expansão 306 e um ganho de zero para a

estimativas de potência abaixo. A curva de ganho final é tomada como o mínimo destas duas curvas constituintes.

O limite de compressão 304, a relação de compressão 303, e o ganho no limite de compressão são parâmetros fixos. Sua escolha determina como o envelope e espectro do sinal da fala são processados em uma faixa particular. Idealmente eles são selecionados de acordo com uma fórmula prescritiva que determina relações de ganho e compressão apropriadas em faixas respectivas para um grupo de ouvintes dada a sua acuidade de audição. Um exemplo de tal fórmula prescritiva é NAL-NLI, que foi desenvolvida pelo Laboratório Nacional de Acústica, Austrália, e é descrita por H. Dillon em "Prescribing hearing aid performance" [H. Dillon (Ed.), Hearing Aids (pp. 249-261); Sydney; Boomerang Press, 2001.] No entanto, eles podem ser também baseados simplesmente na preferência de ouvinte. O limite de compressão 304 e a relação de compressão 303 em uma faixa particular pode adicionalmente depender de parâmetros específicos para um dado programa de audio, tal como o nível médio de diálogo em uma trilha sonora de filme.

Considerando que o limite de compressão pode ser fixo, o limite de expansão 306 é preferencialmente adaptável e varia em resposta ao sinal de entrada. O limite de expansão pode assumir qualquer valor dentro da faixa dinâmica do sistema, incluindo valores maiores do que o limite de compressão. Quando o sinal de entrada é dominado por fala, um sinal de controle descrito abaixo aciona o limite de expansão em direção a níveis baixos de forma que o nível de entrada é mais alto do que a faixa de estimativas de potências para as quais a expansão é aplicada (veja Figuras 3a e 3b). Naquela condição, os ganhos aplicados ao sinal são dominados pela característica de compressão do processador. A Figura 3b apresenta um exemplo de função de ganho representando tal condição.

Quando o sinal de entrada é dominado por audio diferente da fala, o sinal de controle aciona o limite de expansão para níveis altos de forma que o nível de entrada tende a ser mais baixo do que o limite de expansão. Naquela condição a maioria dos componentes do sinal não recebe nenhum ganho. A Figura 3c apresenta um exemplo de função de ganho repre-

sentando tal condição.

As estimativas de potência de faixa da discussão precedente podem ser derivadas analisando as saídas de um banco de filtro ou a saída de uma transformação de domínio de tempo-para-frequência, tal como a

5 DFT (transformada de Fourier discreta), MDCT (transformada de coseno discreta modificada) ou transformadas de wavelet. As estimativas de potência podem ser também substituídas por medidas que são relacionadas à força de sinal tal como o valor absoluto médio do sinal, a energia de Teager, ou por medidas perceptivas tais como sonoridade. Adicionalmente, as estimati-

10 vas da potência de faixa podem ser suavizadas no tempo para controlar a taxa na qual ganho muda.

De acordo com um aspecto da invenção, o limite de expansão é idealmente colocado de tal forma que quando o sinal é de fala o nível de sinal está acima da região expansiva da função de ganho e quando o sinal é

15 de audio diferente da fala o nível de sinal está abaixo da região expansiva da função de ganho. Como é explicado abaixo, isto pode ser obtido monitorando o nível de audio de não-fala e colocando o limite de expansão em relação àquele nível.

Certos monitores de nível da técnica anterior fixam um limite abaixo do qual a expansão descendente (ou silenciador ("squelch")) é aplicada como parte de um sistema de redução de ruído que busca discriminar entre ruído de audio desejável e indesejável. Veja, por exemplo, as Patentes

20 US 3803357, 5263091,

5774557, e 6005953. Em contraste, aspectos da presente invenção requerem diferenciar entre fala por um lado e todos os restantes sinais de audio, como música e efeitos, no outro. O ruído monitorado na técnica anterior é caracterizado por envelopes temporais e espectrais que flutuam muito menos do que aqueles de audios desejáveis. Adicionalmente, o ruído tem frequentemente formas espectrais distintivas que são conhecidas *a priori*.

25

30 Tais características diferenciantes são exploradas por monitores de ruído na técnica anterior. Em contraste, aspectos da presente invenção monitoram o nível de sinais de audio de não-fala. Em muitos casos, tais sinais de audio

de não-fala exibem variações em seu envelope e forma espectral que são pelo menos tão grandes quanto aqueles de sinais de audio da fala. Consequentemente, um monitor de nível empregado na presente invenção requer analisar características de sinal adequadas para a distinção entre audio de
5 fala e não-fala ao invés de entre fala e ruído.

A Figura 4 mostra como o ganho de aperfeiçoamento da fala em uma faixa de frequência pode ser derivado da estimativa de potência de sinal daquela faixa. Referindo-se agora à Figura 4, uma representação de um sinal de uma faixa limitada 401 é passada para um estimador de potência ou
10 dispositivo de estimação ("Estimativa de Potência") 402 que gera uma estimativa de potência de sinal 403 naquela faixa de frequência. Aquela estimativa de potência de sinal é passada para uma transformação de potência para ganho ou função de transformação ("Curva de Ganho") 404, que pode ser da forma do exemplo ilustrado nas Figuras 3a-c. A transformação de po-
15 tência para-ganho ou função de transformação 404 gera um ganho de faixa 405 que pode ser usado para modificar a potência de sinal na faixa (não-mostrado).

A estimativa de potência de sinal 403 é também passada para um dispositivo ou função ("Monitor de Nível") 406 que monitora o nível de
20 todos os componentes de sinal na faixa que são de não-fala. O monitor de nível 406 pode incluir um circuito ou função de manter mínimo vazamento ("Manter Mínimo") 407 com uma taxa de vazamento adaptável. Esta taxa de vazamento é controlada por uma constante de tempo 408 que tende a ser baixa quando a potência do sinal é dominada por fala e alta quando a potên-
25 coa do sinal é dominado por audio diferente da fala. A constante de tempo 408 pode ser derivada de informação contida na estimativa de potência de sinal 403 na faixa. Especificamente, a constante de tempo pode ser monotonicamente relacionada à energia do envelope de sinal de faixa na faixa de frequência entre 4 e 8 Hz. Aquela característica pode ser extraída por um
30 filtro de banda de passagem apropriadamente sintonizado ou função de filtragem ("Banda de passagem") 409.

A saída da Banda de passagem 409 pode ser relacionada à

constante de tempo 408 por uma função de transferência ("Constante de Potência-para-Tempo") 410. A estimativa de nível dos componentes de não-fala 411, que é gerado pelo Monitorador de Nível 406, é a entrada para uma transformação ou função de transformação ("Limite de Potência-para-Expansão") 412 que relaciona a estimativa do nível de fundo para um limite de expansão 414. A combinação do monitorador de nível 406, transformação 412, e expansão descendente (caracterizada pela relação de expansão 305) corresponde ao VAD 108 das Figuras 1a e 1b.

A transformação 412 pode ser uma simples adição, isto é, o limite de expansão 306 pode ser um número fixo de decibéis acima do nível estimado do áudio de não-fala 411. Alternativamente, a transformação 412 que relaciona o nível de fundo estimado 411 ao limite de expansão 306 pode depender de uma estimativa independente da probabilidade do sinal de banda que se fala 413. Deste modo, quando a estimativa 413 indica uma probabilidade alta do sinal ser falada, o limite de expansão 306 é abaixado. Reciprocamente, quando a estimativa 413 indicar uma probabilidade baixa do sinal ser fala, o limite de expansão 306 é aumentado. A estimativa de probabilidade de fala 413 pode ser derivada de uma única característica de sinal ou de uma combinação de características de sinal que distinguem fala de outros sinais. Ela corresponde à saída 109 da SVO 107 nas FIGS 1a e 1b.

Características de sinal adequados e métodos de processá-los para derivar uma estimativa de probabilidade de fala 413 são conhecidas por aqueles qualificados na técnica. Exemplos são descritos nas Patentes US 6,785,645 e 6,570,991, bem como também no pedido de patente 20040044525, e nas referências contidas nelas.

Incorporação por Referência

As seguintes patentes, pedidos e publicações de patentes, são por este meio incorporadas por referência, cada uma em sua totalidade.

Patente dos Estados Unidos 3,803,357; Sacks, 9 de abril de 1974, "Noise Filter".

Patente dos Estados Unidos 5,263,091; Waller, Jr., 16 de novembro de 1993, "Intelligent automatic threshold circuit".

Patente dos Estados Unidos 5,388,185; Terry, e outros, 7 de fevereiro de 1995, "System for adaptive processing of telephone voice signals".

Patente dos Estados Unidos 5,539,806; Allen, e outros, 23 de julho de 1996, "Method for customer selection of telephone sound enhancement".

Patente dos Estados Unidos 5,774,557; Slater, 30 de junho de 1998, "Autotracking microphone squelch for aircraft intercom systems".

Patente dos Estados Unidos 6,005,953; Stuhlfelner, 21 de dezembro de 1999, "Circuit arrangement for improving the signal-to-noise ratio".

Patente dos Estados Unidos 6,061,431; Knappe, e outros, 9 de maio de 2000, "Method for hearing loss compensation in telephony systems based on telephone number resolution".

Patente dos Estados Unidos 6,570,991; Scheirer, e outros, 27 de maio de 2003, "Multi-feature speech/music discrimination system".

Patente dos Estados Unidos 6,785,645; Khalil, e outros, 31 de agosto de 2004, "Real-time speech and music classifier".

Patente dos Estados Unidos 6,914,988; Irwan, e outros, July 5, 2005, "Áudio reproducing device".

Pedido de Patente Publicado US 2004/0044525; Vinton, Mark Stuart, e outros, 4 de março de 2004 "Controlling loudness of speech in signals that contain speech and other types of audio material".

"Dynamic Range Control via Metadata" by Charles Q. Robinson and Kenneth Gundry, Convention Paper 5028, 107th Audio Engineering Society Convention, New York, 24-27 de setembro de 1999.

Implementação

A invenção pode ser implementada em hardware ou software, ou uma combinação de ambos (por exemplo, conjuntos lógicos programáveis). A menos que especificado em contrário, os algoritmos incluídos como parte da invenção não são inerentemente relacionados a qualquer computador ou outro aparelho em particular. Em particular, várias máquinas de propósito geral podem ser usadas com programas escritos de acordo com o preceitos

deste, ou pode ser mais conveniente construir aparelhos mais especializados (por exemplo, circuitos integrados) para executar os passos requeridos pelo método. Deste modo, a invenção pode ser implementada em um ou mais programas de computador executando em um ou mais sistemas de computador programáveis, cada um compreendendo pelo menos um processador, pelo menos um sistema de armazenamento de dados (incluindo memória volátil e não volátil e/ou elementos de armazenamento), pelo menos um dispositivo ou porta de entrada, e pelo menos um dispositivo ou porta de saída. O código de programa é aplicado aos dados de entrada executar as funções descritas aqui e gerar informação de saída. A informação de saída é aplicada a um ou mais dispositivos de saída, em um modo conhecido.

Cada um de tais programas pode ser implementado em qualquer linguagem de computador desejada (incluindo, de máquina, de montagem, ou procedimento de alto nível, lógica, ou linguagens de programação orientadas a objeto) para comunicar com um sistema de computador. Em qualquer caso, a linguagem pode ser uma linguagem compilada ou interpretada.

Cada um de tais programas de computador é preferencialmente armazenado em um meio ou dispositivo de armazenamento ou carregado nele (por exemplo, memória de estado sólido, ou meio magnético ou ótico) legível por um computador programável de propósito geral ou especial, para configurar e operar o computador quando o meio ou dispositivo de armazenamento é lido pelo sistema de computador para executar os procedimentos descritos aqui. O sistema inventivo pode ser também considerado para ser implementado como um meio de armazenamento legível por computador, configurado com um programa de computador, onde o meio de armazenamento assim configurado causa um sistema de computador a operar em uma maneira específica e predefinida para executar as funções descritas aqui.

Foram descritas um grande numero de versões da invenção. Não obstante, será entendido que várias modificações possam ser feitas sem se afastar do espírito e escopo da invenção. Por exemplo, alguns dos

passos descritos aqui podem ser independente de ordem, e deste modo podem ser executados em uma ordem diferente daquela descrita.

REIVINDICAÇÕES

1. Método para aperfeiçoar a fala em audio de entretenimento, compreendendo processar, em resposta a um ou mais controles, o dito audio de entretenimento para aperfeiçoar a clareza e inteligibilidade de partes da fala do audio de entretenimento, o dito processamento incluindo
- 5 variar o nível do audio de entretenimento em cada uma das faixas de frequência múltipla de acordo com uma característica de ganho que relaciona o nível de sinal de faixa ao ganho, e
- gerar um controle para variar a dita característica de ganho em
- 10 cada faixa de frequência, o dito gerador incluindo
- caracterizar segmentos de tempo do dito áudio de entretenimento como (a) fala ou não-fala ou (b) como provável para ser fala ou não-fala, em que as dita caracterizações eperam em uma única faixa de frequência banda larga,
- 15 obter, em cada uma das faixas da dita frequência múltipla, uma medida de flutuações em níveis da fala,
- monitorar, em cada uma das faixas da dita frequência múltipla, o mínimo do nível de áudio na faixa, o tempo de resposta do monitoramento respondendo à dita medida de flutuações nos níveis da fala,
- 20 transformar o mínimo monitorado em cada faixa em um correspondente nível de limite adaptativo, e
- desviar cada dito correspondente nível de limite adaptativo com o resultado da dita caracterização para produzir o dito controle para cada faixa.
- 25 2. Método para aperfeiçoar a fala em audio de entretenimento, compreendendo processar, em resposta a um ou mais controles, o dito audio de entretenimento para aperfeiçoar a clareza e inteligibilidade de partes da fala do audio de entretenimento, o dito processamento incluindo variar o nível do audio de entretenimento em cada uma das faixas de frequência múltipla, de acordo com uma característica de ganho que relaciona nível de sinal de faixa a ganho, e gerar um controle para variar a dita característica de ganho em cada faixa de frequência, o dito gerador incluindo receber caracte-
- 30

rizações de segmentos de tempo do dito audio de entretenimento como (a) fala ou não-fala ou (b) como provável para ser fala ou não-fala, em que as dita caracterizações referem-se a uma única faixa de frequência de banda larga, obter, em cada uma das faixas da dita frequência múltipla, uma
5 medida de flutuações em níveis da fala, monitorar, em cada uma das faixas da dita frequência múltipla, o mínimo do nível de audio na faixa, o tempo de resposta do monitoramento respondendo à dita medida de flutuações nos níveis da fala, transformar o minimo monitorado em cada faixa em um correspondente nível de limite adaptativo, e desviar cada dito correspondente
10 nível de limite adaptativo com o resultado da dita caracterização para produzir o dito controle para cada faixa.

3. Método de acordo com a reivindicação 1 ou 2, em que existe acesso a uma evolução de tempo do audio de entretenimento antes e depois de um ponto de processamento, e em que a dita geração de um controle
15 responde a pelo menos algum audio depois do ponto de processamento.

4. Método de acordo com qualquer uma das reivindicações 1 a 3 em que o dito processamento opera de acordo com um ou mais parâmetros de processamento.

5. Método de acordo com a reivindicação 4, em que o ajuste de
20 um ou mais parâmetros é em resposta ao audio de entretenimento de tal forma que uma métrica inteligibilidade da fala do audio processado é tanto maximizado quanto impulsionado acima de um nível de limite desejado.

6. Método de acordo com a reivindicação 5, em que o audio de entretenimento compreende múltiplos canais de audio nos quais um canal é
25 principalmente de fala e o ou mais outros canais são principalmente de não-fala, em que a métrica da inteligibilidade da fala é baseada no nível do canal da fala e o nível no um ou mais outros canais.

7. Método de acordo com a reivindicação 5 ou 6, em que a métrica de inteligibilidade da fala é também baseada no nível de ruído em um
30 ambiente de escutar em que o audio processado é reproduzido.

8. Método de acordo com qualquer uma das reivindicações 4 a 7, em que o ajuste de um ou mais parâmetros é em resposta a um ou mais

descritores de longo prazo do audio de entretenimento.

9. Método de acordo com a reivindicação 8, em que um descritor de termo longo é o nível médio de diálogo do áudio de entretenimento.

5 10. Método de acordo com a reivindicação 8 ou 9, em que um descriptor de longo termo é uma estimativa de processamento já aplicada ao áudio de entretenimento.

10 11. Método de acordo com a reivindicação 4, em que o ajuste de um ou mais parâmetros está de acordo com uma fórmula prescritiva, em que a fórmula prescritiva relaciona a acuidade de audição de um ouvinte ou grupo de ouvintes ao um ou mais parâmetros.

12. Método de acordo com a reivindicação 4, em que o ajuste de um ou mais parâmetros está de acordo com as preferências de um ou mais ouvintes.

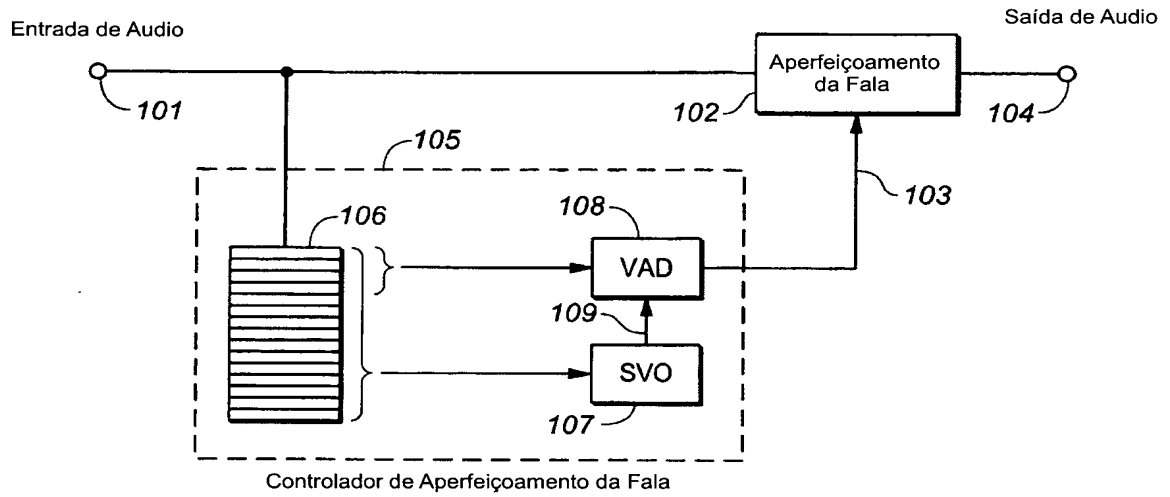
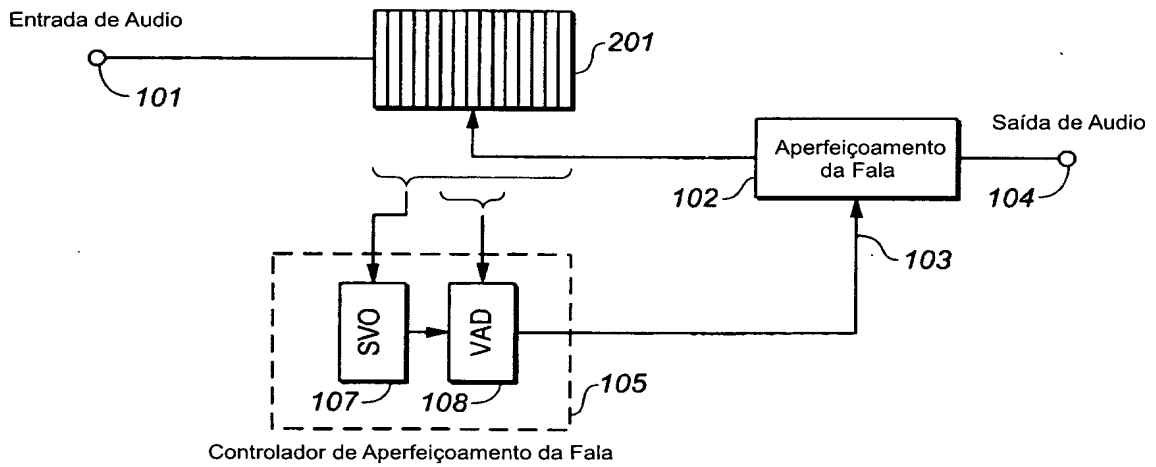
15 13. Método de acordo com qualquer uma das reivindicações 1 a 12, em que o dito processamento provê controle dinâmico de faixa, equalização dinâmica, estreitamento espectral, extração da fala, redução de ruído, ou outra ação de aperfeiçoar a fala.

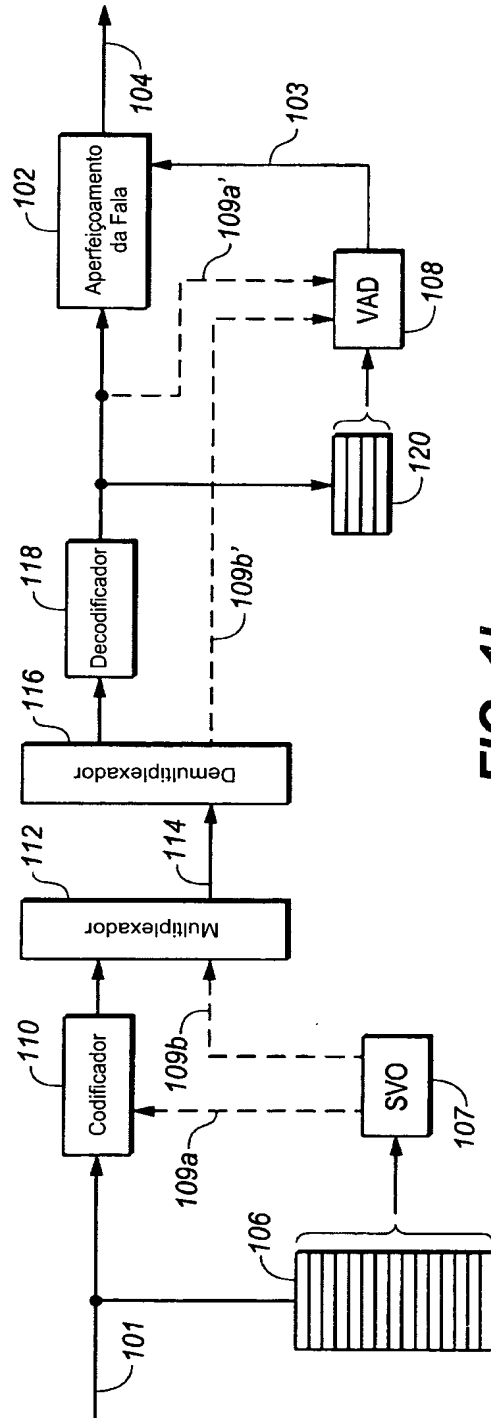
20 14. Método de acordo com a reivindicação 13, em que o controle dinâmico de faixa é provido por uma função de compressão/expansão de faixa.

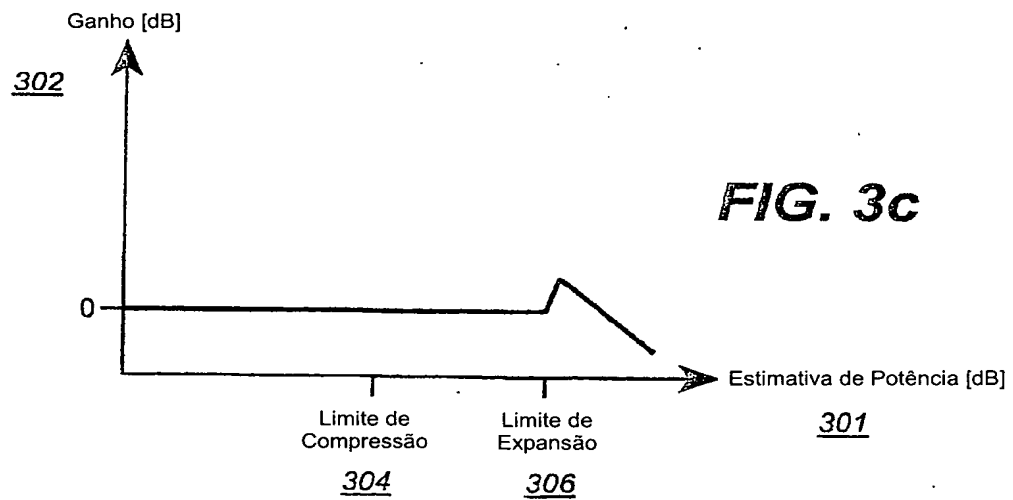
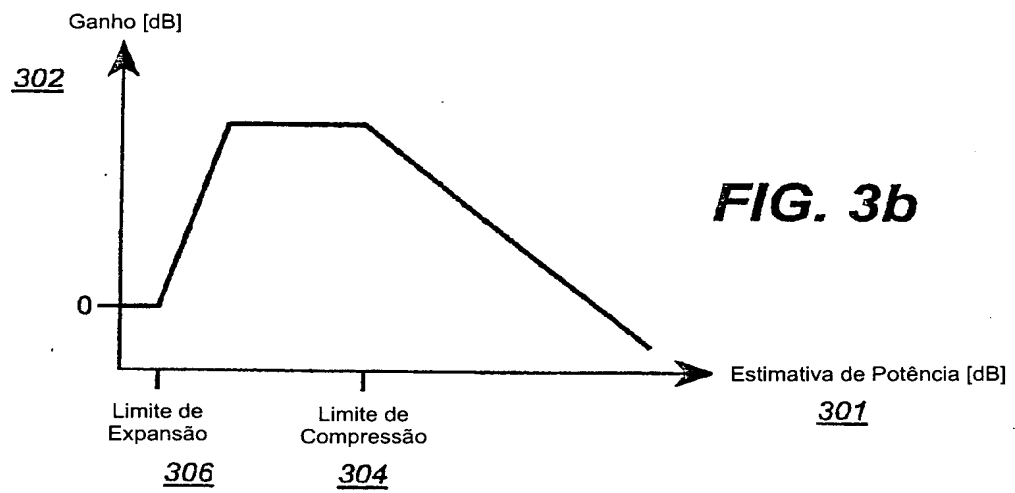
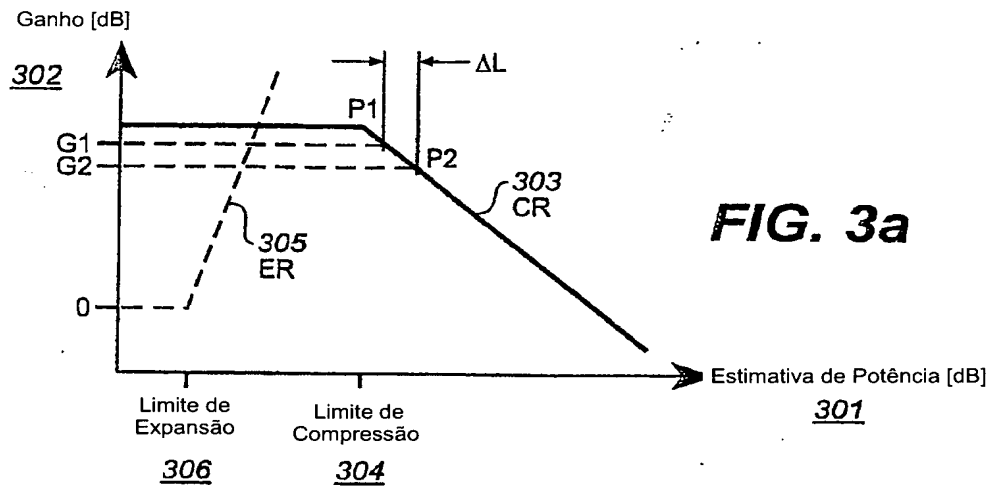
15. Aparelho compreendendo meios adaptados para executar o método como definido em qualquer uma das reivindicações 1 a 14.

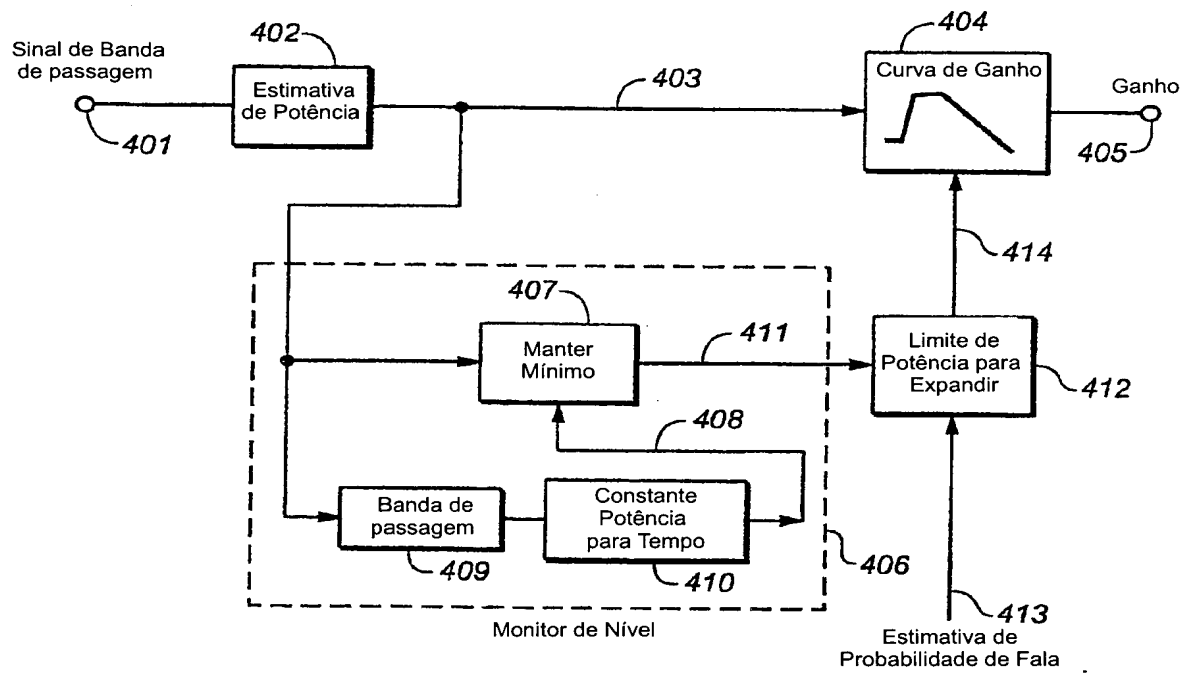
25 16. Programa de computador, armazenado em um meio legível de computador para causar a um computador executar o método como definido em qualquer uma das reivindicações 1 a 14.

17. Meio legível de computador armazenando nele o programa de computador executando o método como definido em qualquer uma das reivindicações 1 a 14.

**FIG. 1a****FIG. 2**

**FIG. 1b**



**FIG. 4**

RESUMO

Patente de Invenção: **"APERFEIÇOAMENTO DA FALA EM AUDIO DE ENTRETENIMENTO"**.

5 A presente invenção refere-se a processamento de sinal de audio. Mais especificamente, a invenção refere-se a aperfeiçoamento de audio de entretenimento, como audio de televisão, aperfeiçoar a clareza e inteligibilidade da fala, como diálogo e narrativa de audios. A invenção refere-se a métodos, aparelho para executar tais métodos, e a software armazenado em um meio legível de computador para causar a um computador a executar
10 tais métodos.