



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria
e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(11) PI 0112901-5 B1

(22) Data do Depósito: 26/07/2001

(45) Data de Concessão: 22/12/2015
(RPI 2346)



(54) Título: MÉTODOS PARA RECONHECER UMA AMOSTRA DE ÁUDIO, E, SISTEMA DE COMPUTADOR PARA REALIZAR O MESMO

(51) Int.Cl.: G10L 17/00

(30) Prioridade Unionista: 31/07/2000 US 60/222,023, 20/04/2001 US 09/839476

(73) Titular(es): LANDMARK DIGITAL SERVICES LLC

(72) Inventor(es): AVERY LI-CHUN WANG, JULIUS O. SMITH III

“MÉTODO PARA RECONHECER UMA AMOSTRA DE ÁUDIO, E, SISTEMA DE COMPUTADOR PARA REALIZAR O MESMO”

CAMPO DA INVENÇÃO

A presente invenção diz, em geral, a respeito de recuperação de
5 informação com base em conteúdo. Mais particularmente, a presente invenção diz respeito a reconhecimento de um sinal de áudio, incluindo som ou música, o qual seja altamente distorcido ou contém um alto nível de ruído.

FUNDAMENTOS DA TÉCNICA

Há uma necessidade crescente por reconhecimento automático
10 de música ou de outros sinais de áudio gerados a partir de uma variedade de fontes. Por exemplo, donos de obras protegidas pela lei de direito autoral ou anunciantes estão interessados em obter dados na frequência de difusão de seu material. Serviços de rastreamento de músicas fornecem listas de reprodução das maiores estações de rádio nos grandes mercados. Consumidores gostariam
15 de identificar músicas ou propagandas difundidas no rádio, de tal maneira que eles pudessem comprar produtos musicais ou outros produtos e serviços novos e interessantes. Qualquer espécie de reconhecimento de som sob demanda ou contínuo é ineficiente e trabalhoso quando é realizado por seres humanos. Um método automatizado de reconhecimento de música ou som forneceria, desse
20 modo, um benefício significativo aos consumidores, artistas e a uma variedade de indústrias. À medida que o paradigma de distribuição de músicas muda de compras em lojas para tê-las baixadas através da Internet, é perfeitamente viável ligar o reconhecimento de música implementado por computador diretamente com a compra pela Internet ou outros serviços à base da Internet.

25 Tradicionalmente, reconhecimento de músicas tocadas na rádio era realizado pela coincidência de estações de rádio e de horários nos quais as músicas eram tocadas com listas de reprodução providas ou pelas estações de rádio ou a partir de fontes de terceiros. Tal método é limitado inerentemente a apenas as estações de rádio para as quais a informação encontra-se disponível.

Outros métodos baseiam-se no embutimento de códigos inaudíveis dentro dos sinais radiodifundidos. Os sinais embutidos são decodificados no receptor para extrair informação de identificação sobre o sinal radiodifundido. A desvantagem deste método é que os dispositivos de decodificação especiais são necessários para identificação de sinais, e apenas aquelas músicas com códigos embutidos podem ser identificadas.

Qualquer reconhecimento de áudio em larga escala exige algum tipo de recuperação de áudio com base em conteúdo, na qual um sinal radiodifundido não identificado é comparado com uma base de dados de tais sinais conhecidos para identificar os sinais da base de dados similares ou idênticos. Notar que a recuperação de áudio com base em conteúdo é diferente de uma recuperação de áudio existente pelos mecanismos de busca na rede, nos quais apenas o texto de metadados cercando ou associado com os arquivos de áudio é buscado. Além disso, vale notar que enquanto reconhecimento de fala é útil para a conversão de sinais de voz em texto que pode, a seguir, ser indexado e buscado usando técnicas bem conhecidas, não é aplicável à grande maioria de sinais de áudio que contêm músicas e sons. De alguma forma, recuperação da informação de áudio é análoga à recuperação de informação com base em texto provida pelos mecanismos de busca. De outra forma, contudo, reconhecimento de áudio não é análogo: sinais de áudio não apresentam entidades facilmente identificáveis, tais como palavras que fornecem identificadores para a busca e indexação. Como tal, esquemas de recuperação de áudio corrente indexam os sinais de áudio por características de percepção computadas que representam várias qualidades ou características do sinal.

Uma recuperação de áudio com base em conteúdo é tipicamente efetuada analisando-se um sinal de consulta “query” para obter um número de características representativas, e posteriormente, aplicando-se uma medida de similaridade com as características derivadas para a localização dos arquivos de base de dados que são mais similares ao referido sinal de consulta “query”.

A similaridade dos objetos recebidos é necessariamente um reflexo daquelas características de percepção selecionadas. Numerosos métodos de recuperação com base em conteúdo estão disponíveis na técnica. Por exemplo, patente US No. 5.210.820, expedida para Kenyon, revela um método de reconhecimento de sinal, no qual sinais recebidos são processados e amostrados para se obter valores de sinal em cada ponto de amostragem. Os momentos estatísticos dos valores amostrados são então computados para gerar um vetor de característica que pode ser comparado com os identificadores dos sinais armazenados para recuperar sinais similares. As patentes US Nos. 4.450.531 e 4.843.562, ambas expedidas para Kenyon et al., revelam métodos de classificação de informação radiodifundida similares, em que correlações cruzadas são computadas entre os sinais não identificados e os sinais de referência armazenados.

Um sistema para a recuperação de documentos de áudio pela similaridade acústica é descrito em J. T. Foote, “Content-Based Retrieval of Music and Audio”, em C.-C. J. Kuo et al., editor, *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, volume 3.229, páginas 138 a 147, 1997. Vetores de característica são calculados por parametrização de cada arquivo de áudio em coeficientes “cepstral” de escala “mel”, e uma árvore de quantização é criada a partir dos dados de parametrização. Para realizar uma consulta, um sinal desconhecido é parametrizado para se obter vetores de característica que são então ordenadas em nós de folha da árvore. Um histograma é coletado para cada nó de folha, gerando assim um vetor N-dimensional que representa o sinal desconhecido. A distância entre dois destes vetores é indicativa da similaridade entre dois arquivos de som. No método em questão, o esquema de quantização supervisionado aprende a distinguir características de áudio, enquanto ignorando variações não importantes, com base em classes às quais os dados de treinamento estão designados por um ser humano. Dependendo do sistema de classificação, diferentes características acústicas são escolhidas como importantes. Assim, tal método é mais adequado para encontrar similaridades entre músicas e ordenar

músicas em classes do que ele é para reconhecer música.

Um método para a análise, armazenamento, recuperação, bem como segmentação, com base em conteúdo, de informação de áudio é descrito na patente US No. 5.918.223, expedida para Blum et al. Nesse método, uma
5 série de características acústicas, tais como sonoridade, grave, passo, brilho, largura de banda, e coeficientes “cepstral” de frequência “Mel”, são medidos em intervalos de tempo periódicos de cada arquivo. As medições estatísticas daquelas características são recolhidas e combinadas para formar um vetor de característica. Os arquivos de dados de áudio dentro de uma base de dados são
10 recuperados com base na similaridade de seus vetores de característica para o vetor de característica de um arquivo não identificado.

Um problema ‘chave’ de todos os métodos de reconhecimento de áudio acima da técnica anterior é que os mesmos tendem a falhar quando os sinais a serem reconhecidos são sujeitos a distorção linear e não linear causadas
15 por, por exemplo, ruído de fundo, erros e quedas de transmissão, interferência, filtragem limitada de banda, quantização, deformação no tempo, e compressão digital de qualidade de voz. Naqueles métodos da técnica anterior, quando uma amostra de som distorcida é processada para se obter características acústicas, só uma fração das características derivadas para a gravação original são achadas.
20 O vetor de característica resultante é, portanto, não muito similar ao vetor de característica da gravação original, e é improvável que reconhecimento correto possa ser realizado. Continua a permanecer uma necessidade de um sistema de reconhecimento de som que se comporte bem sob condições de ruído e distorção elevados.

25 Um outro problema com os métodos da técnica anterior é que os mesmos exigem muita computação e não escalonam bem. Reconhecimento em tempo real é, desta maneira, não possível utilizando os métodos da técnica anterior com bases de dados enormes. Em tais sistemas, é inviável se ter uma base de dados que possua mais de poucas centenas ou milhares de gravações.

O tempo de busca nos métodos da técnica anterior tendem a crescer de forma linear com o tamanho da base de dados, tornando o escalonamento de milhões de gravações de músicas economicamente inviável. Tais métodos de Kenyon também necessitam de grandes bancos de hardware de processamento de sinal digital especializados.

Métodos comerciais existentes têm frequentemente exigências rígidas para a amostra de entrada, para que a mesma seja capaz de executar o reconhecimento. Por exemplo, eles exigem que a música inteira ou pelo menos 30 segundos da música sejam amostrados ou, ainda, exigem que a música seja amostrada desde o início. Eles também apresentam dificuldade em reconhecer múltiplas músicas misturadas uma com a outra em uma única corrente. Todas essas desvantagens tornam os métodos da técnica anterior inviáveis para uso em muitas aplicações práticas.

OBJETIVOS E VANTAGENS

Por conseguinte, é um objetivo principal da presente invenção o de fornecer um método de reconhecimento de um sinal de áudio sujeito a um alto nível de ruído e distorção.

É um objetivo adicional da invenção o de fornecer um método de reconhecimento que possa ser realizado em tempo real com base em apenas uns poucos segundos do sinal a ser identificado.

É um outro objetivo da invenção o de fornecer um método de reconhecimento que possa reconhecer músicas com base em amostras de quase qualquer lugar dentro da música, não apenas do começo.

É um objetivo adicional da invenção o de fornecer um método de reconhecimento que não requeira que amostras de música sejam codificadas ou correlacionadas com listas de música ou estações de rádio em particular.

É um objetivo adicional da invenção o de fornecer um método de reconhecimento que possa reconhecer cada uma das múltiplas gravações de música misturadas uma com a outra em uma única corrente.

É um outro objetivo da invenção o de fornecer um sistema de reconhecimento de música no qual a música desconhecida possa ser fornecida para o sistema a partir de qualquer ambiente por qualquer método virtualmente conhecido.

5

SUMÁRIO

Esses objetivos e vantagens são alcançados por um método de reconhecimento de uma amostra de mídia, tal como uma amostra de áudio, de acordo com um índice de base de dados de um grande número de arquivos de mídia conhecidos. Tal índice de base de dados contém impressões digitais que representam características em localizações particulares dos arquivos de mídia indexados. A amostra de mídia desconhecida é identificada com um arquivo de mídia na base de dados (o arquivo de mídia vencedor), cujas localizações relativas de impressões digitais mais próximas coincidem com as localizações relativas das impressões digitais da amostra. No caso de arquivos de áudio, a evolução no tempo de impressões digitais do arquivo vencedor coincide com a evolução no tempo das impressões digitais na amostra.

O método é preferencialmente implementado num sistema de computador distribuído e contém as seguintes etapas: determinar um conjunto de impressões digitais em localizações particulares da amostra; localizar as referidas impressões digitais que coincidem no índice de base de dados; gerar correspondências entre as localizações na amostra e as localizações no arquivo possuindo impressões digitais equivalentes; e identificar arquivos de mídia para os quais um número significativo das correspondências são substancialmente relacionadas linearmente. O arquivo tendo o maior número de correspondências relacionadas linearmente é tido como o arquivo de mídia vencedor. Um método de identificação de arquivos com um grande número de correspondências é o de realizar o equivalente à varredura para uma linha diagonal no diagrama de dispersão gerado a partir dos pares de correspondências. Numa modalidade, a identificação dos arquivos de mídia com um grande número de correspondências

lineares envolve a busca de apenas um primeiro subconjunto dos arquivos de mídia. Os arquivos no primeiro subconjunto têm uma probabilidade maior de serem identificados do que os arquivos que não estão no primeiro subconjunto. A probabilidade de identificação é preferencialmente baseada em frequência empírica ou medidas de recência de identificações anteriores, juntamente com projeções *a priori* de frequência de identificação. Se nenhum arquivo de mídia for identificado no primeiro subconjunto, então o segundo subconjunto, o qual contém o resto dos arquivos, é buscado. Alternativamente, os arquivos podem ser classificados por probabilidade e buscados na ordem da classificação. Tal busca é encerrada quando um arquivo é localizado.

De preferência, as localizações particulares dentro das amostras são computadas de maneira a serem reproduzíveis na dependência da amostra. Tais localizações computáveis reproduzivelmente são chamadas de “marcos”. As ditas impressões digitais são, preferencialmente, valores numéricos. Numa modalidade, cada impressão digital representa um número de características da amostra de mídia em cada localização, ou desviam ligeiramente da localização.

O método é, particularmente, útil para reconhecer amostras de áudio, em cujo o caso as localizações particulares são pontos de tempo dentro da amostra de áudio. Tais pontos de tempo ocorrem, por exemplo, na máxima local de normas L_p espectrais da amostra de áudio. Impressões digitais podem ser computadas por qualquer análise da amostra de áudio, e são, de preferência, invariáveis a prolongamento de tempo da amostra. Os exemplos de impressões digitais incluem impressões digitais de fatia espectral, impressões digitais de múltiplas fatias, coeficientes LPC, coeficientes “cepstral”, e componentes de frequência de picos de espectrograma.

A presente invenção também provê um sistema para implementar o método acima, contendo um: objeto de marco para computar as localizações particulares, objeto de impressão digital para computar as impressões digitais, índice de base de dados que contém as localizações do arquivo e as impressões

digitais para os arquivos de mídia, e objeto de análise. O dito objeto de análise implementa o método ao localizar impressões digitais coincidentes no índice de base de dados, gerar correspondências, e analisar as correspondências para selecionar o arquivo de mídia vencedor.

5 Também está fornecido um dispositivo de armazenamento de programa acessível por um computador, consubstanciando, de modo tangível, um programa de instruções executáveis pelo computador para realizar etapas do método para o método acima.

Adicionalmente, a invenção provê um método de criação de um
10 índice de um número de arquivos de áudio em uma base de dados, contendo as seguintes etapas: computar um conjunto de impressões digitais em localizações particulares de cada arquivo; e armazenar as impressões digitais, localizações e identificadores dos arquivos em uma memória. Uma impressão digital, uma localização, e um identificador correspondentes estão associados na memória
15 na forma de uma tríplice. Preferencialmente, as localizações, que podem ser pontos de tempo dentro do arquivo de áudio, são computadas na dependência do arquivo e são reproduzíveis. Por exemplo, aqueles pontos de tempo podem ocorrer numa máxima local de normas L_p espectrais do arquivo de áudio. Em alguns casos, cada impressão digital, que é de preferência um valor numérico,
20 representa um número de características daquele arquivo perto da localização particular. As impressões digitais podem ser computadas a partir de qualquer análise ou processamento de sinal digital do arquivo de áudio. Exemplos das impressões digitais incluem impressões digitais de fatia espectral, impressões digitais de múltiplas fatias, coeficientes LPC, coeficientes “cepstral”, como
25 também os componentes de frequência de picos de espectrograma, e os picos de espectrograma conectados.

Finalmente, a invenção propicia métodos para identificação de amostras de áudio incorporando impressões digitais invariáveis a prolongamento de tempo e várias buscas hierárquicas.

BREVE DESCRIÇÃO DAS FIGURAS

Figura 1 é um fluxograma de um método da invenção para o reconhecimento de uma amostra de som;

Figura 2 é um diagrama de bloco de um sistema de computador distribuído exemplar para implementar o método da Figura 1;

Figura 3 é um fluxograma de um método de construção de um índice de base de dados de arquivos de som utilizado no método da Figura 1;

Figura 4 ilustra, de maneira esquemática, marcos e impressões digitais computados para uma amostra de som;

Figura 5 é um gráfico de normas L4 para uma amostra de som, ilustrando a seleção de marcos;

Figura 6 é um fluxograma de uma modalidade alternativa para construir um índice de base de dados de arquivos de som utilizados no método da Figura 1;

Figuras 7A, 7B e 7C ilustram um espectrograma com pontos salientes e pontos salientes conectados indicados;

Figuras 8A, 8B e 8C ilustram conjuntos de índice, uma lista de índice, e uma lista de índice principal do método da Figura 3;

Figuras 9A, 9B e 9C ilustram uma lista de índice, uma lista de candidatos, uma lista de dispersão do método da Figura 1;

Figuras 10A e 10B são diagramas de dispersão que ilustram a identificação correta e falta de identificação, respectivamente, de uma amostra de som desconhecida.

DESCRIÇÃO DETALHADA

A presente invenção fornece um método para reconhecimento de uma amostra de mídia exógena, de acordo com uma base de dados contendo um grande número de arquivos de mídia conhecidos. Ela também fornece um método para gerar um índice de base de dados, que possibilita a busca eficiente utilizando o método de reconhecimento da invenção. Enquanto que a discussão

a seguir se refere, basicamente, a dados de áudio, deve-se compreender que o método da presente invenção pode ser aplicado a qualquer tipo de amostras de mídia e arquivos de mídia, incluindo, mas não limitado a, texto, áudio, vídeo, imagem e quaisquer combinações de multimídia de tipos de mídias individuais.

- 5 No caso de áudio, a presente invenção é particularmente útil para reconhecer amostras que contenham altos níveis de distorção linear e não linear causados, por exemplo, por ruído de fundo, erros e quedas de transmissão, interferência, filtragem limitada de banda, quantização, deformação no tempo, e compressão digital da qualidade de voz. Como será aparente a partir da descrição abaixo,
- 10 a invenção funciona sob tais condições, já que pode reconhecer corretamente um sinal distorcido mesmo se apenas uma pequena fração das características computadas sobreviverem à distorção. Qualquer tipo de áudio, incluindo som, voz, música ou ainda combinações de tipos, pode ser reconhecido pela presente invenção. Exemplos de amostras de áudio incluem música gravada, programas
- 15 de radiodifusão, e propagandas.

- Como usada aqui, uma amostra de mídia exógena constitui um segmento de dado de mídia de qualquer tamanho obtido de uma variedade de fontes como descrito abaixo. Para o reconhecimento ser executado, a amostra deve ser uma capitulação de parte de um arquivo de mídia indexada em uma
- 20 base de dados usada pela presente invenção. O arquivo de mídia indexado pode ser tido como uma gravação original, e a amostra como uma versão distorcida e/ou abreviada ou uma capitulação da gravação original. Tipicamente, a amostra corresponde a apenas uma pequena parte do arquivo indexado. Por exemplo, o reconhecimento pode ser executado num segmento de dez segundos de uma
- 25 música com cinco minutos indexada na base de dados. Muito embora o termo “arquivo” seja usado para descrever a entidade indexada, a entidade pode estar em qualquer formato para o qual os valores necessários (infra descritos) podem ser obtidos. Adicionalmente, não há nenhuma necessidade de se armazenar ou ter acesso ao arquivo depois que os valores são obtidos.

Um diagrama de blocos que ilustra, conceitualmente, as etapas gerais de um método 10 da presente invenção é exibido na figura 1. As etapas individuais são descritas em maiores detalhes abaixo. O método identifica um arquivo de mídia vencedor, um arquivo de mídia cujas localizações relativas de impressões digitais características mais próximas coincidem com aquelas localizações relativas das mesmas impressões digitais da dita amostra exógena. Após uma amostra exógena ser capturada na etapa 12, marcos e impressões digitais são computados na etapa 14. Os ditos marcos ocorrem em localizações particulares, por exemplo, pontos de tempo, dentro da amostra. A localização dentro da amostra dos marcos é preferencialmente determinada pela amostra propriamente dita, isto é, é dependente das qualidades da amostra, e pode ser reproduzida. A saber, os mesmos marcos são computados para o mesmo sinal toda vez que o processo é repetido. Para cada marco, uma impressão digital caracterizando uma ou mais características da amostra em ou perto do marco é obtida. A proximidade de uma característica com um marco é definida pelo método de impressão digital utilizado. Em alguns casos, uma característica é considerada próxima de um marco se ela corresponder claramente ao marco e não a um marco anterior ou subsequente. Em outros casos, tais características correspondem a múltiplos marcos adjacentes. Por exemplo, impressões digitais de texto podem ser sequências de palavras, impressões digitais de áudio podem ser componentes espectrais, e impressões digitais de imagem podem ser valores de pixel RGB. Duas modalidades gerais da etapa 14 estão descritas abaixo, uma na qual os marcos e impressões digitais são computados sequencialmente, e uma na qual eles são computados simultaneamente.

Na etapa 16, as impressões digitais de amostras são utilizadas para recuperar conjuntos de impressões digitais coincidentes armazenadas em um índice de base de dados 18, em que as impressões digitais coincidentes estão associadas com marcos e identificadores de um conjunto de arquivos de mídia. O conjunto de valores de identificadores e marcos de arquivo recuperados são,

então, usados para gerar pares de correspondência (etapa 20) contendo marcos de amostra (computados na etapa 14) e marcos de arquivo recuperados, em que as mesmas impressões digitais foram computadas. Os pares de correspondência resultantes são, em seguida, ordenados por um identificador de música, gerando

5 conjuntos de correspondências entre os marcos de amostra e marcos de arquivo para cada arquivo aplicável. Cada conjunto é varrido para o alinhamento entre os marcos de arquivo e marcos de amostra. Ou seja, correspondências lineares nos pares de marcos são identificadas, e o conjunto é marcado de acordo com o número de pares que estão relacionados linearmente. Uma correspondência

10 linear ocorre quando um grande número de correspondentes localizações de amostras e localizações de arquivos puderem ser descritas, substancialmente, com a mesma equação linear, dentro de uma tolerância permitida. Por exemplo, se inclinações de um número de equações descrevendo um conjunto de pares de correspondência variarem em $\pm 5\%$, então o conjunto de correspondências

15 na íntegra é considerado como estando relacionado linearmente. Naturalmente, qualquer tolerância adequada pode ser selecionada. O identificador do conjunto com a marcação mais alta, a saber, com o maior número de correspondências relacionadas linearmente, é o identificador de arquivo vencedor, que se encontra localizado e retornado na etapa 22.

20 Como descrito adicionalmente abaixo, reconhecimento pode ser realizado com um componente de tempo proporcional ao logaritmo do número de entradas na base de dados. Reconhecimento pode ser feito, essencialmente, em tempo real, mesmo com uma base de dados muito grande. *I.e.*, uma amostra pode ser reconhecida, à medida que ela está sendo obtida, com um retardo de

25 tempo pequeno. O método pode identificar um som com base em segmentos de 5-10 segundos e mesmo tão baixo quanto de 1-3 segundos. Numa modalidade preferida, a análise dos marcos e impressões digitais, etapa 14, é realizada em tempo real, à medida que a amostra está sendo capturada na etapa 12. Consultas em base de dados (etapa 16) são realizadas, à medida que as impressões digitais

de amostras se tornam disponíveis, e os resultados de correspondências estão acumulados e são varridos, periodicamente, para as correspondências lineares. Assim, todas as etapas do método ocorrem de forma simultânea, e não do modo linear sequencial como sugerido na figura 1. Note-se que o método é em parte

5 análogo a um mecanismo de busca de texto: usuário submete uma amostra de consulta, e arquivo coincidente indexado na base de dados de som é retornado.

O referido método é tipicamente implementado como software que roda em um sistema de computador, com etapas individuais que são mais eficaz/eficientemente implementadas como módulos de software independentes.

10 Consequentemente, um sistema que implementa a presente invenção pode ser considerado como consistindo em um objeto de marco e de impressão digital, uma base de dados indexada, e um objeto de análise para a busca do índice de base de dados, correspondências de computação, e a identificação do arquivo vencedor. No caso de marco e impressão digital sequencial, o objeto de marco e

15 impressão digital pode ser considerado como sendo objetos distintos de marco e impressão digital. Código de instrução de computador para aqueles objetos diferentes é armazenado em uma memória de um ou mais computadores e é executado por um ou mais processadores do computador. Numa modalidade, os objetos de código estão agrupados em um único sistema de computador, tal

20 como um computador pessoal com base em Intel ou outra estação de trabalho. Numa modalidade preferencial, tal método é implementado por intermédio de um agrupamento “cluster” conectado em redes de unidades de processamento central (CPUs), nas quais diferentes objetos de software são executados por diferentes processadores com o propósito de distribuir a carga computacional.

25 Alternativamente, cada uma das CPUs pode ter uma cópia de todos os objetos de software, permitindo uma rede homogênea de elementos configurados de forma idêntica. Nessa última configuração, cada CPU tem um subconjunto do índice de base de dados e é responsável pela busca de seu próprio subconjunto de arquivos de mídia.

A despeito da invenção não ser limitada a qualquer sistema de hardware particular, um exemplo de uma modalidade preferida de um sistema de computador distribuído 30 é ilustrado de forma esquemática na figura 2. O sistema 30 contém um grupo de processadores com base em Linux 32a a 32f

5 conectado por uma arquitetura de barramento de multiprocessamento 34 ou por um protocolo de rede tal como o protocolo de computação de grupo Beowulf, ou uma mistura dos dois. Num arranjo deste tipo, o índice de base de dados é, preferencialmente, armazenado na memória de acesso randômico (RAM) em pelo menos um nó 32a no grupo, garantindo que a busca de impressão digital

10 ocorra muito rapidamente. Os nós de computação que correspondem a outros objetos, tais como nós de marco 32c e 32f, nós de impressão digital 32b e 32e, e nó de varredura de alinhamento 32d, não exigem uma RAM tão grande como o nó ou nós 32a suportando o índice de base de dados. O número dos nós de computação designados para cada objeto pode ser, desta forma, escalonado de

15 acordo com a necessidade, de tal forma que nenhum objeto único se torne um afunilamento. A rede computacional é, por conseguinte, altamente paralelizável e pode processar, adicionalmente, múltiplas consultas de reconhecimento de sinal simultâneas sendo distribuídas entre recursos computacionais disponíveis. Note-se que isso torna possíveis as aplicações nas quais grandes números de

20 usuários podem solicitar o reconhecimento e receber resultados quase que em tempo real.

Numa modalidade alternativa, determinados objetos dos objetos funcionais são acoplados de modo mais justo, enquanto permanecem acoplados de maneira menos justa a outros objetos. Por exemplo, o objeto de marco e de

25 impressão digital pode residir numa localização fisicamente separada do resto dos objetos computacionais. Um exemplo disto é a associação justa dos objetos de marco e impressão digital com o processo de captura de sinal. Neste arranjo, o objeto de marco e de impressão digital pode ser incorporado como hardware adicional ou software adicional embutido, por exemplo, num telefone móvel,

num navegador de Protocolo de Aplicação Sem Fio (WAP), assistente digital pessoal (PDA), ou outro terminal remoto, tal como terminal de cliente de um mecanismo de busca de áudio. Em um serviço de busca de áudio com base na Internet, tal como um serviço de identificação de conteúdo, o objeto de marco e impressão digital pode ser incorporado no aplicativo de navegador do cliente como um conjunto conectado de instruções de software ou um módulo *plug-in* de software, tal como uma biblioteca de ligação dinâmica da Microsoft (DLL). Nestas modalidades, a captura de sinal combinado, objeto de marco e impressão digital, constitui o terminal de cliente do serviço. O terminal de cliente envia um sumário extraído de característica da amostra de sinal capturado contendo pares de marcos e impressões digitais para o terminal de servidor, que realiza o reconhecimento. Envio desse sumário extraído de característica ao servidor, ao invés do sinal capturado cru, é vantajoso visto que a quantidade de dados é muito reduzida, frequentemente por um fator de 500 ou mais. Tal informação pode ser enviada em tempo real através de um canal lateral de baixa largura de banda juntamente com ou ao invés de, por exemplo, uma corrente de áudio transmitida ao servidor. Isso permite a realização da invenção através de redes de comunicações públicas, as quais oferecem larguras de banda de dimensões relativamente pequenas para cada usuário.

O método será, agora, descrito em detalhes, com referência às amostras de áudio e arquivos de áudio indexados numa base de dados de som. O método consiste de dois componentes amplos, construção de índice de base de dados de som e reconhecimento de amostra.

Construção de Índice de Base de Dados

Antes do reconhecimento de som poder ser efetuado, um índice de base de dados de som, o qual pode ser buscado, deve ser construído. Como usada aqui, uma base de dados é qualquer coleção indexada de dados, e não está limitada a bases de dados comercialmente disponíveis. No índice de base de dados, elementos relacionados de dados são associados uns com os outros,

e elementos individuais podem ser usados para recuperar os dados associados. O índice de base de dados de som contém um conjunto de índices para cada arquivo ou gravação na coleção selecionada ou biblioteca de gravações, que podem incluir fala, música, propagandas, assinaturas de som, ou outros sons.

- 5 Cada gravação também possui um identificador único, `sound_ID`. A base de dados de som propriamente dita não armazena necessariamente os arquivos de áudio para cada gravação, mas `sound_ID` pode ser utilizado para recuperar os arquivos de áudio de outro lugar. Espera-se que o índice da base de dados de som seja muito grande, contendo índices para milhões ou, até mesmo, bilhões
- 10 de arquivos. As novas gravações são de preferência acrescentadas de maneira incremental ao índice de base de dados.

Um diagrama em blocos de um método preferencial 40 para a construção do índice de base de dados de som que pode ser buscado de acordo com uma primeira modalidade é ilustrado na figura 3. Nessa modalidade, os

15 marcos são primeiro computados, e então impressões digitais são computadas em ou perto dos marcos. Como será evidente para aquele versado na técnica, os métodos alternativos podem ser concebidos para a construção do índice de base de dados. Em particular, muitas das etapas listadas abaixo são opcionais, mas servem para a geração de um índice de base de dados que é mais eficientemente

20 buscada. Enquanto a eficiência de busca é importante para reconhecimento de som em tempo real a partir de grandes bases de dados, pequenas bases de dados podem ser buscadas de modo relativamente rápido mesmo se elas não tiverem sido ordenadas de forma otimizada.

Para indexar a base de dados de som, cada gravação na coleção

25 é submetida a uma análise de marco e impressão digital que gera um conjunto de índice para cada arquivo de áudio. A figura 4 ilustra de forma esquemática um segmento de uma gravação de som para o qual marcos (LM) e impressões digitais (FP) foram computados. Aqueles marcos ocorrem em pontos de tempo específicos do som e têm valores em desvio de unidades de tempo a partir do

começo do arquivo, enquanto aquelas impressões digitais caracterizam o som em ou perto de um marco em particular. Assim, nessa modalidade, cada marco para um arquivo em particular é singular, enquanto a mesma impressão digital pode ocorrer várias vezes dentro de um arquivo único ou múltiplos arquivos.

5 Na etapa 42, cada gravação de som é marcada usando métodos para encontrar localizações distintivas e reproduzíveis dentro da gravação de som. Um algoritmo de marco preferido é capaz de marcar os mesmos pontos de tempo dentro de uma gravação de som, independentemente da presença de ruído e outra distorção linear e não linear. Alguns dos métodos de marcação são
10 conceitualmente independentes do processo de impressão digital infra descrito, porém pode ser escolhido para otimizar o desempenho do último. A marcação resulta numa lista de pontos de tempo $\{\text{landmark}_k\}$ dentro da gravação de som em que impressões digitais são subsequentemente calculadas. Um bom esquema de marcação marca cerca de 5-10 marcos por segundo de gravação de som; é
15 óbvio que a densidade de marcação depende da quantidade de atividade dentro da gravação de som.

 Uma variedade de técnicas é possível para computar os marcos, todas das quais estão dentro do escopo da presente invenção. Estes processos técnicos específicos utilizados para implementar os esquemas de marcação da
20 invenção são conhecidos na técnica e não serão discutidos em detalhes. Uma técnica de marcação simples, conhecida como Norma de Energia, é calcular a potência instantânea a cada ponto de tempo possível na gravação e selecionar a máxima local. Uma forma de se fazer isso é calcular o envelope ao retificar e filtrar a forma de onda diretamente. Outro jeito é calcular a transformada de
25 Hilbert (quadratura) do sinal e empregar a soma das magnitudes quadradas da transformada de Hilbert e o sinal original.

 O método da Norma de Energia de marcação é ótimo para se encontrar transientes no sinal de som. Norma de Energia é, na realidade, um caso especial daquela Norma L_p Espectral mais geral na qual $p = 2$. A Norma

Lp Espectral geral é calculada toda vez ao longo do sinal de som pelo cálculo de um espectro de tempo curto, por exemplo, por meio de uma Transformada Rápida de Fourier (FFT) de Janela Hanning. Uma modalidade preferida utiliza uma taxa de amostragem de 8000 Hz, um tamanho de quadro de FFT de 1024 amostras, e um passo de 64 amostras para cada fatia de tempo. Norma Lp para cada fatia de tempo é então calculada como a soma da $p^{\text{ésima}}$ energia dos valores absolutos dos componentes espectrais, opcionalmente seguida pela retirada da raiz da $p^{\text{ésima}}$. Como antes, os marcos são escolhidos como a máxima local dos valores resultantes ao longo do tempo. Um exemplo do método da Norma Lp Espectral é ilustrado na figura 5, um gráfico da norma L4 como uma função do tempo para um sinal de som em particular. As linhas tracejadas na máxima local indicam a localização dos marcos escolhidos.

Quando $p=\infty$, a norma L^∞ é efetivamente a norma máxima. A saber, o valor da norma é o valor absoluto do componente espectral maior na fatia espectral. Essa norma resulta em marcos robustos e bom desempenho de reconhecimento geral, e é preferida para música tonal.

Alternativamente, marcos espectrais de “múltiplas fatias” podem ser calculados pela retirada da soma de $p^{\text{ésima}}$ energias de valores absolutos de componentes espectrais através de múltiplas fatias de tempo em desvios fixos ou variáveis um do outro, ao invés de uma única fatia. O encontro da máxima local dessa soma estendida permite a otimização da colocação das impressões digitais de múltiplas fatias, descritas abaixo.

Uma vez que marcos tenham sido computados, uma impressão digital é computada em cada ponto de tempo do marco na gravação na etapa 44. A impressão digital é geralmente um valor ou um conjunto de valores que resume um conjunto de características na gravação em ou perto do ponto de tempo. Numa modalidade atualmente preferida, cada impressão digital é um valor numérico único que é uma função “hashed” de múltiplas características. Tipos possíveis de impressões digitais incluem as impressões digitais de fatia

espectral, impressões digitais de múltiplas fatias, coeficientes LPC, bem como coeficientes “cepstral”. Naturalmente, qualquer tipo de impressão digital que caracteriza o sinal ou as características do sinal perto de um marco está dentro do escopo da presente invenção. Impressões digitais podem ser computadas por qualquer tipo de processamento de sinal digital ou análise de frequência do sinal.

5 Para gerar impressões digitais de fatia espectral, uma análise de frequência é realizada nas proximidades de cada ponto de tempo de marco para extrair vários picos espectrais do topo. Um valor simples de impressão digital é apenas o valor de frequência único do pico espectral mais forte. O uso de um
10 pico simples como este resulta num reconhecimento surpreendentemente bom na presença de ruído; não obstante, as impressões digitais de fatia espectral de frequência única tendem a gerar mais esquemas falso positivos do que outros esquemas de impressão digital visto que eles não são singulares. O número de falso positivos pode ser reduzido usando-se impressões digitais consistindo em
15 uma função dos dois ou três picos espectrais mais fortes. Entretanto, pode haver uma suscetibilidade mais elevada a ruído se o segundo pico espectral mais forte não for suficientemente forte o bastante para distingui-lo de seus competidores na presença de ruído. Em outros termos, o valor da impressão digital calculada pode não ser suficientemente robusto para ser reproduzível de maneira confiável.
20 Independentemente disso, o desempenho desse caso também é bom.

Para levar vantagem da evolução no tempo de muitos sons, um conjunto de intervalos de tempo é determinado acrescentando-se um conjunto de desvios de tempo para um ponto de tempo do marco. Em cada intervalo de tempo resultante, a impressão digital de fatia espectral é calculada. O conjunto
25 resultante da informação de impressão digital é então combinado para formar uma impressão digital de múltiplos tons ou de múltiplas fatias. Cada impressão digital de múltiplas fatias é muito mais singular do que a impressão digital da fatia espectral única, pois a mesma rastreia evolução temporal, resultando em menos falsa coincidências na busca de índice de base de dados descrita abaixo.

Experimentos indicam que devido à sua singularidade aumentada, impressões digitais de múltiplas fatias computadas a partir do pico espectral singular mais forte em cada um dentre os dois intervalos de tempo resultam em computação muito mais rápida (cerca de 100 vezes mais rápida) na busca de índice de base de dados subsequente, porém com alguma degradação numa porcentagem de reconhecimento na presença de ruído significativo.

Alternativamente, ao invés de utilizar um desvio fixo ou desvios fixos a partir de um dado intervalo de tempo, para calcular uma impressão digital de múltiplas fatias, desvios variáveis podem ser usados. O desvio variável para a fatia escolhida é o desvio para o próximo marco, ou um marco em uma certa faixa de desvio a partir do marco de “âncora” para a impressão digital. Nesse caso, a diferença de tempo entre os marcos também é codificada na impressão digital, juntamente com a informação de múltiplas frequências. Pela adição de mais dimensões às impressões digitais, as mesmas se tornam singulares e têm uma menor chance de falsa coincidência.

Além de componentes espectrais, outras características espectrais podem ser extraídas e usadas como impressões digitais. Análise de codificação preditiva linear (LPC) extrai as características linearmente previsíveis de um sinal, tais como picos espectrais, bem como a forma espectral. LPC é bastante conhecida na arte de processamento de sinal digital. Para a presente invenção, coeficientes LPC de fatias de forma de onda ancorados nas posições de marco podem ser utilizados como impressões digitais por “hashing” dos coeficientes LPC quantizados em um valor de índice.

Aqueles coeficientes “cepstral” são úteis como uma medida de periodicidade e podem ser usados para caracterizar sinais que são harmônicos, tais como vozes ou muitos instrumentos musicais. A análise “cepstral” é bem conhecida na técnica de processamento de sinal digital. Para fins da presente invenção, um número de coeficientes “cepstral” são “hashed” juntos num índice e utilizados como uma impressão digital.

Uma modalidade alternativa 50, em que os marcos e impressões digitais são computados simultaneamente, é mostrada na figura 6. As etapas 42 e 44 da figura 3 são substituídas pelas etapas 52, 54 e 56. Como descrito abaixo, uma função multidimensional é computada a partir da gravação de som na etapa 52, e marcos (54) e impressões digitais (56) são extraídos da função.

Em uma implementação da modalidade da figura 6, os marcos e impressões digitais são computados a partir de um espectrograma da gravação do som. Um espectrograma é uma análise de tempo-frequência de uma gravação de som na qual quadros em janelas e quadros sobrepostos de amostras de som são analisados espectralmente, tipicamente usando uma Transformada Rápida de Fourier (FFT). Como antes, uma modalidade preferida utiliza uma taxa de amostragem de 8000 Hz, um tamanho de quadro de FFT de 1024 amostras, e um passo de 64 amostras para cada intervalo de tempo. Um exemplo de um espectrograma é ilustrado na figura 7A. O tempo está no eixo horizontal, e a frequência está no eixo vertical. Cada quadro de FFT sequencial é empilhado verticalmente em intervalos espaçados de forma homogênea correspondentes ao longo do eixo do tempo. Uma representação do espectrograma demonstra a densidade de energia em cada ponto de tempo-frequência; áreas mais escuras no gráfico representam uma maior densidade de energia. Espectrogramas são bem conhecidos na técnica de processamento de sinal de áudio. Para a presente invenção, marcos e impressões digitais podem ser obtidos de pontos salientes tais como na máxima local do espectrograma, circulada no espectrograma da figura 7B. Por exemplo, coordenadas de tempo e de frequência de cada pico são obtidas, o tempo considerado como sendo o marco, e a frequência utilizada para a computação da impressão digital correspondente. Este marco de pico de espectrograma é semelhante à norma L^∞ , na qual o valor máximo absoluto da norma determina a localização do marco. No espectrograma, todavia, a busca máxima local ocorre através de partes do plano de tempo-frequência, ao invés de através de todo o intervalo de tempo.

Nesse contexto, o conjunto dos pontos salientes que resulta da análise de extração de ponto de uma gravação de som é referido como sendo uma constelação. Para uma constelação que consiste numa máxima local, uma análise preferida é a de seleccionar pontos que são a energia máxima do plano tempo-frequência através de uma vizinhança ao redor de cada ponto selecionado. Por exemplo, um ponto na coordenada (t_o, f_o) é selecionado se for o ponto de energia máxima dentro de um retângulo com os cantos (t_o-T, f_o-F) , (t_o-T, f_o+F) , (t_o+T, f_o-F) e (t_o+T, f_o+F) i.e. um retângulo com lados de comprimento $2T$ e $2F$, com T e F escolhidos para proporcionar um número apropriado de pontos de constelação. As junções do retângulo também poderão variar em tamanho de acordo com o valor de frequência. Obviamente, qualquer forma da região pode ser usada. O critério de energia máxima também pode ser ponderado tal que um pico de energia de tempo-frequência competidor seja inversamente ponderado de acordo com uma métrica de distância no plano tempo-frequência, ou seja, pontos mais distantes têm menor ponderação. Por exemplo, a energia pode ser ponderada como

$$\frac{S(t, f)}{1 + C_t(t - t_o)^2 + C_f(f - f_o)^2},$$

onde $S(t, f)$ é o valor ao quadrado da magnitude do espectrograma no ponto (t, f) , e C_t e C_f são valores positivos (não necessariamente constantes). Outras funções de distância-ponderação são possíveis. As restrições de seleção máxima local podem ser aplicadas a outros esquemas (a saber, não máximos) de extração de característica de ponto saliente, e estão dentro do escopo da invenção.

Tal método resulta em pares de valores que são muito similares à impressão digital espectral de frequência única descrita acima, com muitas das mesmas propriedades. O método tempo-frequência do espectrograma gera mais pares de marco/impressão digital do que o método de frequência singular, mas também poderá resultar em muitas falsa coincidências naquele estágio de coincidência descrito abaixo. Contudo, o mesmo provê marcação e impressão digital mais robustos do que a impressão digital espectral de frequência única,

já que ruído dominante na amostra de som pode não se estender para todas as partes do espectro em cada fatia. Isto é, existem, muito provavelmente, alguns pares de marco e impressão digital em partes do espectro que não são afetados pelo ruído dominante.

5 Esse método de marcação e impressão digital do espectrograma é um caso especial dos métodos de análise das características que computam uma função multidimensional do sinal do som, no qual uma das dimensões é o tempo, e localizam pontos salientes nos valores de função. Os pontos salientes podem ser a máxima local, a mínima local, cruzamentos pelo ‘zero’, ou outras
10 características distintivas. Os marcos são tomados como sendo as coordenadas de tempo dos pontos salientes, e já as impressões digitais correspondentes são computadas a partir de pelo menos uma das coordenadas remanescentes. Por exemplo, a(s) coordenada(s) de não tempo do ponto saliente multidimensional podem ser “hashed” juntas para formar uma impressão digital funcional multi-
15 dimensional.

O método de desvio variável descrito acima para as impressões digitais espectrais de múltiplas fatias pode ser aplicado ao espectrograma ou a outras impressões digitais de função multidimensional. Nesse caso, os pontos numa constelação são conectados entre si para formar os pontos conectados,
20 como ilustrado no espectrograma exibido na figura 7C. Cada ponto na referida constelação serve como um ponto de ancoragem definindo o tempo de marco, e os valores de coordenadas restantes dos outros pontos são combinados para formar a impressão digital conectada. Os pontos que estão mais pertos um do outro, por exemplo, como definidos abaixo, são conectados juntos para formar
25 impressões digitais de característica agregada mais complexas que podem ser mais facilmente distinguidas e buscadas. Como ocorre com aquelas impressões digitais espectrais de múltiplas fatias, o objetivo de combinar uma informação a partir de múltiplos pontos salientes conectados numa única impressão digital é o de criar uma maior diversidade dos valores de impressão digital possíveis,

com isto diminuindo a probabilidade de falsa coincidência, isto é, diminuindo a probabilidade de que a mesma impressão digital descreva duas amostras de música diferentes.

Em princípio, cada um dos pontos salientes N pode ser conectado a cada outro ponto, num esquema de conexão de dois pontos, produzindo cerca de $N^2/2$ combinações. Similarmente, para uma conexão de ponto K , o número de combinações possíveis resultantes de uma constelação é da ordem de N^K . A fim de se evitar uma explosão combinatória como esta, é desejável restringir a vizinhança dos pontos que estão conectados juntos. Um modo de realizar uma restrição como esta é definir uma “zona alvo” para cada ponto de ancoragem. Um ponto de ancoragem é, então, conectado com pontos em sua zona alvo. É possível seleccionar um subconjunto de pontos dentro da zona alvo para a sua conexão - nem todo ponto precisa ser conectado. Por exemplo, apenas os pontos associados com os picos mais fortes na zona alvo podem ser conectados. Uma zona alvo pode ter uma forma fixa ou variar de acordo com características do ponto de ancoragem. Um exemplo simples de uma zona alvo de um ponto de ancoragem (t_o, f_o) para uma constelação de pico de espectrograma é o conjunto de pontos (t, f) na faixa do espectrograma, de tal forma que t esteja no intervalo $[t_o + L, t_o + L + W]$, onde L é o trajeto para o futuro e W é a largura da zona alvo. Neste esquema, todas as frequências são permitidas na zona alvo. L ou W podem ser variáveis, por exemplo, se um mecanismo de controle de taxa for usado para modular o número de combinações de conexão sendo produzidas. De maneira alternativa, restrições de frequência podem ser implementadas, por exemplo, por restrição da zona alvo, tal que a frequência f esteja no intervalo $[f_o - F, f_o + F]$, onde F é um parâmetro de limitação. Uma das vantagens de uma restrição de frequência é que, em psico-acústica, é sabido que melodias tendem a concordar melhor quando sequências de notas têm frequências que estão próximas entre si. Uma restrição deste tipo pode possibilitar mais desempenho de reconhecimento “psico-acusticamente realístico”, muito embora modelagem psico-acústica não

seja necessariamente um objetivo da invenção. Também é possível considerar a regra oposta, em que f é escolhido do lado de fora da região $[f_o-F, f_o+F]$. Isso força a conexão dos pontos que são diferentes uns dos outros em frequência, possivelmente evitando casos nos quais os artefatos de extração de constelação produzam sequências de obturação de pontos de tempo-frequência que estão próximos em tempo e tenham a mesma frequência. Como ocorre com os outros parâmetros de localidade, F não é necessariamente uma constante e pode, por exemplo, ser uma função de f_o .

Quando incluindo coordenadas de tempo de pontos salientes de não ancoragem nos valores de impressão digital, os valores de tempo relativos devem ser usados para permitir que as impressões digitais sejam invariáveis no tempo. Por exemplo, dita impressão digital pode ser uma função de (i) valores de coordenada de não tempo e/ou (ii) diferença(s) dos valores de coordenada de tempo correspondentes dos pontos salientes. A(s) diferença(s) de tempo podem ser consideradas, por exemplo, com relação ao ponto de ancoragem ou como diferenças sucessivas entre pontos salientes sequenciais no conjunto conectado. Os valores de coordenadas e diferenças podem ser empacotados em campos de bit concatenados para formar a impressão digital “hashed”. Como será aparente aqueles versados na técnica, muitas outras formas de conjuntos de mapeamento de valores de coordenadas em um valor de impressão digital existem e estão dentro do escopo da presente invenção.

Uma instanciação concreta desse esquema utiliza os picos de espectrograma conectados $N>1$ com coordenadas (t_k, f_k) , $k=1, \dots, N$. Então, (i) o tempo t_1 do primeiro pico é tomado como o tempo do marco, e (ii) as diferenças de tempo $\Delta t_k = t_k - t_1$, $k=2, \dots, N$, mais as frequências f_k , $k=1, \dots, N$, dos picos conectados são “hashed” juntos para formar um valor de impressão digital. A dita impressão digital pode ser computada a partir de todos ou a partir de um subconjunto de todas as coordenadas disponíveis Δt_k e f_k . Por exemplo, algumas ou todas as coordenadas de diferença de tempo podem ser omitidas se desejado.

Outra vantagem da utilização de múltiplos pontos para formar a impressão digital é que codificação da impressão digital pode ser feita invariável com relação a prolongamento de tempo, por exemplo, quando uma gravação de som é reproduzida numa velocidade diferente daquela velocidade de gravação original. Essa vantagem se aplica a ambos os métodos de espectrograma e de intervalo de tempo. Note-se que num sinal de tempo prolongado, diferenças de tempo e frequência possuem uma relação alternada (por exemplo, diminuindo a diferença de tempo entre dois pontos por um fator de dois dobros da frequência). Esse método aproveita a vantagem desse fato pela combinação de diferenças de tempo e frequências de um modo que remova o prolongamento de tempo a partir da impressão digital.

Por exemplo, em um caso de pico de espectrograma de ponto N com os valores de coordenadas (t_k, f_k) , $k=1, \dots, N$, os valores intermediários disponíveis a serem “hashed” numa impressão digital são $\Delta t_k = t_k - t_1$, $k=2, \dots, N$, e f_k , $k=1, \dots, N$. Os valores intermediários podem então se tornarem invariáveis com relação ao prolongamento do tempo considerando-se uma das frequências como uma frequência de referência, digamos f_1 , e formando (i) quocientes com as frequências remanescentes e (ii) produtos com as diferenças de tempo.

Por exemplo, valores intermediários podem ser $g_k = f_k/f_1$, $k=2, \dots, N$, e $s_k = \Delta t_k f_1$, $k=2, \dots, N$. Se a amostra for acelerada por um fator de α , então a frequência f_k se torna αf_k e a diferença de tempo Δt_k se torna $\Delta t_k/\alpha$, de tal modo que $g_k = \alpha f_k/\alpha f_1 = f_k/f_1$, e, ainda, $s_k = (\Delta t_k/\alpha) (\alpha f_1) = \Delta t_k f_1$. Esses novos valores intermediários são então combinados usando uma função para formar um valor de impressão digital “hashed” que é independente do prolongamento de tempo. Por exemplo, os valores g_k e s_k podem ser “hashed” ao empacotá-los em campos de bits concatenados.

Alternativamente, ao invés de uma frequência de referência, uma diferença de tempo de referência pode ser utilizada, por exemplo, Δt_2 . Neste caso, os novos valores intermediários são calculados como (i) quocientes

$\Delta t_k/\Delta t_2$ com as diferenças de tempo remanescentes e (ii) produtos $\Delta t_2 f_k$ com as frequências. Este caso é equivalente ao uso de uma frequência de referência, visto que os valores resultantes podem ser formados a partir dos produtos e quocientes dos valores g_k e s_k acima. As alternâncias das razões de frequência

5 podem ser empregadas igualmente de forma eficiente; somas e diferenças dos valores logarítmicos dos valores intermediários originais também podem ser substituídos pelos produtos e diferenças, respectivamente. Qualquer valor de impressão digital independente de prolongamento de tempo obtido por aquelas comutações, substituições e permutas de operações matemáticas se encontram

10 dentro do escopo da invenção. Adicionalmente, as múltiplas frequências de referência ou as diferenças de tempo de referência, que também relativizam as diferenças de tempo, podem ser utilizadas. O uso de múltiplas frequências de referência ou diferenças de tempo de referência é equivalente ao uso de uma única referência, pois o mesmo resultado pode ser alcançado por manipulação

15 aritmética dos valores g_k e s_k .

Retornando às figuras 3 e 6, análises de marcação e impressão digital por qualquer um dos métodos acima resultam num conjunto de índices para cada sound_ID, conforme ilustrado na figura 8A. Um conjunto de índices para uma dada gravação de som é uma lista de pares de valores (*i.e.*, impressão

20 digital, marco). Cada gravação indexada possui tipicamente na ordem de mil pares (impressão digital, marco) em seu conjunto de índices. Naquela primeira modalidade descrita acima, na qual as ditas técnicas de marcação e impressão digital são essencialmente independentes, as mesmas podem ser tratadas como módulos separados e intercambiáveis. Dependendo do sistema, qualidade de

25 sinal, ou tipo de som a ser reconhecido, um dentre um número de módulos de marcação ou de impressão digital diferentes podem ser empregados. De fato, já que o conjunto de índices é constituído simplesmente de pares de valores, é possível e frequentemente preferível utilizar múltiplos esquemas de marcação e impressão digital simultaneamente. Por exemplo, um esquema de marcação

e impressão digital pode ser bom na detecção de padrões de tons únicos, mas ruim na identificação de percussão, ao passo que um algoritmo diferente pode ter os atributos opostos. O uso de múltiplas estratégias de marcação/impressão digital resulta numa faixa mais robusta e mais rica do referido desempenho de reconhecimento. Diferentes técnicas de impressão digital podem ser utilizadas juntas reservando-se determinadas faixas de valores de impressão digital para certos tipos de impressões digitais. Por exemplo, num valor de impressão digital de 32 bits, os primeiros 3 bits podem ser utilizados para especificar quais dos 8 esquemas de impressão digital os seguintes 29 bits estão codificando.

Após conjuntos de índices serem gerados para cada gravação de som a ser indexada na base de dados de som, um índice de base de dados que pode ser buscado é construído de tal maneira a permitir uma busca rápida (isto é, tempo de registro “log”). Isto é efetuado na etapa 46 pela construção de uma lista de tríplexes (impressão digital, marco, sound_ID) obtida por anexação da sound_ID correspondente a cada dueto dentro de cada conjunto de índices. Todas as ditas tríplexes para todas as gravações de som são coletadas em uma grande lista de índice, um exemplo da qual é ilustrado na figura 8B. A fim de otimizar o processo de busca subsequente, a lista de tríplexes é então ordenada de acordo com a impressão digital. Algoritmos de ordenação rápidos são bem conhecidos na técnica e extensivamente discutidos em D.E. Knuth, *The Art of Computer Programming, Vol. 3: Sorting and Searching*, Reading, Massachusetts: Addison-Wesley, 1998, incorporado aqui por referência. Algoritmos de ordenação de alto desempenho podem ser utilizados para ordenar a lista em registro “log” N tempo N , onde N é o número de entradas na lista.

Uma vez que a lista de índice é ordenada, ela é adicionalmente processada na etapa 48 pela segmentação, de tal maneira que cada impressão digital singular na lista seja coletada em uma nova lista de índice principal, um exemplo da qual é ilustrado na figura 8C. Cada entrada na lista de índice principal contém um valor de impressão digital e um apontador para uma lista

de pares (marco, sound_ID). Dependendo do número e caractere das gravações indexadas, uma dada impressão digital pode aparecer centenas de vezes ou mais dentro de toda a coleção. O rearranjo da lista de índice em uma lista de índice principal é opcional, mas economiza memória, já que cada valor de impressão digital aparece apenas uma vez. Isto também acelera a busca de base de dados subsequente, visto que o número efetivo de entradas na lista é muito reduzido para uma lista de valores únicos. Alternativamente, a lista de índice principal pode ser construída pela inserção de cada tríplice em uma árvore B. Existem outras possibilidades de construção da lista de índice principal, como é sabido dos versados na técnica. A lista de índice principal é preferencialmente mantida na memória do sistema, tal como a DRAM, para acesso rápido durante o dito reconhecimento de sinal. A lista de índice principal pode ser mantida naquela memória de um único nó dentro do sistema, conforme ilustrada na figura 2. Alternativamente, a lista de índice principal pode ser fragmentada em pedaços distribuídos entre múltiplos nós computacionais. De preferência, o índice de base de dados de som supra referido é a lista de índice principal ilustrada na figura 8C.

O índice de base de dados de som é, de preferência, construído fora de linha e atualizado de forma incremental, à medida que novos sons são incorporados ao sistema de reconhecimento. Para a atualização da lista, novas impressões digitais podem ser inseridas na localização apropriada na lista de índice principal. Se novas gravações contiverem impressões digitais existentes, os pares correspondentes (marco, sound_ID) são adicionados a listas existentes para aquelas impressões digitais.

Sistema de Reconhecimento

Usando a lista de índice principal gerada como descrita acima, reconhecimento de som é efetuado numa amostra de som exógena, tipicamente, suprida por um usuário interessado na identificação da amostra. Por exemplo, o usuário ouve uma nova música no rádio e gostaria de saber o nome do artista e título da música. A amostra pode surgir a partir de qualquer tipo de ambiente,

tal como, uma radiodifusão, boate, bar, submarino, arquivo de som, segmento de transmissão “streaming” de áudio, ou sistema estéreo, e podem conter ruído de fundo, interrupções ou vozes falando. O usuário pode armazenar a amostra de áudio num dispositivo de armazenagem, tal como uma secretária eletrônica, arquivo de computador, gravador, ou sistema de correio de voz de telefone ou 5 telefone móvel antes de fornecê-la para o sistema para reconhecimento. Com base na configuração do sistema e restrições impostas pelo usuário, a amostra de áudio é fornecida para o sistema de reconhecimento da presente invenção a partir de qualquer número de fontes analógicas ou digitais, tais como, sistema 10 estéreo, televisão, aparelho tocador de CD, radiodifusão, secretária eletrônica, telefone, telefone móvel, difusão de “streaming” na Internet, FTP, arquivo de computador, como um anexo do e-mail, ou qualquer outro meio adequado de transmissão de tal material gravado. Dependendo da fonte, a amostra pode estar na forma de ondas acústicas, ondas de rádio, uma corrente PCM de áudio digital, 15 uma corrente de áudio digital comprimida (tal como Dolby Digital ou MP3) ou uma difusão de “streaming” na Internet. O usuário interage com o sistema de reconhecimento através de uma interface padrão tal como um telefone, telefone móvel, navegador da rede ou e-mail. A amostra pode ser capturada pelo sistema e processada em tempo real, ou ela pode ser reproduzida para processamento a 20 partir de um som capturado previamente (por exemplo, de um arquivo de som). Durante captura, a amostra de áudio é amostrada digitalmente e enviada para o sistema por um dispositivo de amostragem, tal como o microfone. Dependendo do método de captura, a dita amostra fica provavelmente sujeita à degradação adicional devido às limitações do canal ou dispositivo de captura de som.

25 Uma vez que o sinal de som foi convertido em forma digital, o mesmo é processado para reconhecimento. Como ocorre com a construção de conjuntos de índices para arquivos de base de dados, os marcos e impressões digitais são calculados para a amostra usando o mesmo algoritmo que foi usado para o processamento da base de dados de gravação de som. O método opera,

de forma otimizada, se o processamento de uma rendição altamente distorcida de um arquivo de som original resultar em um conjunto idêntico ou similar de pares de marco e impressão digital como fora obtido para a gravação original. O conjunto de índices resultante para a amostra de som é um conjunto de pares

5 de valores analisados, (impressão digital, marco), ilustrados na figura 9A.

Dados os pares para a amostra de som, o índice de base de dados é buscado para localizar, potencialmente, arquivos coincidentes. A busca é feita como a seguir: cada par (impressão digital_k, marco_k) no conjunto de índices da amostra desconhecida é processado pela busca por impressão digital_k na lista

10 de índice principal. Algoritmos de busca rápida numa lista ordenada são bem conhecidos na técnica e extensivamente discutidos em D.E. Knuth, *The Art of Computer Programming, Vol. 3: Sorting and Searching*, Reading, Massachusetts: Addison-Wesley, 1998. Se impressão digital_k for encontrada na lista de índice principal, então sua lista correspondente de par coincidente (marco*_j, sound_ID_j)

15 é copiada e aumentada com o marco_k para formar um conjunto de tríptices da forma (marco_k, marco*_j, sound_ID_j). Nessa notação, um asterisco (*) indica um marco de um dos arquivos indexados na base de dados, ao passo que um marco sem um asterisco se refere à amostra. Em alguns casos, é preferencial que as impressões digitais coincidentes não sejam necessariamente idênticas, mas sim

20 similares; por exemplo, elas poderão diferir dentro de um limiar previamente determinado. Impressões digitais coincidentes, se idênticas ou similares, são referidas como equivalentes. O sound_ID_j na tríptice corresponde ao arquivo tendo o marco com asterisco. Assim, cada tríptice contém dois marcos distintos, um no índice de base de dados e um na amostra, em que as impressões digitais

25 equivalentes foram computadas. Esse processo é repetido para todos k variando pelo conjunto de índices da amostra de entrada. Todas tríptices resultantes são coletadas numa grande lista candidata, exibida na figura 9B. A lista candidata é assim chamada, pois contém sound_IDs dos arquivos de som, que, em virtude de suas impressões digitais coincidentes, são candidatas para identificação com

a amostra de som exógena.

Após a lista candidata ser compilada, a mesma é adicionalmente processada pela segmentação de acordo com sound_ID. Um modo conveniente de se fazer isso é o de ordenar a lista candidata por sound_ID ou inseri-la em uma árvore B. Um grande número de algoritmos de ordenação está disponível na técnica, conforme discutido acima. O resultado deste processo é uma lista de sound_IDs candidatas, cada uma das quais tem uma lista de dispersão de pares de amostra e pontos de tempo do marco de arquivo com o referido sound_ID opcionalmente retirado, $(marco_k, marco*_j)$, como ilustrado na figura 9C. Cada lista de dispersão contém, assim, um conjunto de marcos correspondentes, que correspondem em virtude de eles serem caracterizados por valor de impressão digital equivalente.

A lista de dispersão para cada sound_ID candidata é, a seguir, analisada para determinar se tal sound_ID é uma coincidência para a amostra. Uma etapa de limitação opcional pode ser utilizada primeiro para eliminar um número potencialmente grande de candidatas que possuem listas de dispersão muito pequenas. Claramente, as candidatas possuindo apenas uma entrada em suas listas de dispersão, isto é, apenas uma impressão digital em comum com a amostra, não coincidem com dita amostra. Qualquer número limiar adequado maior do que ou igual a um pode ser utilizado.

Uma vez que aquele número final de candidatas é determinado, a candidata vencedora é localizada. Se o algoritmo seguinte não localiza uma candidata vencedora, então uma mensagem de falha é retornada. Uma perspicácia chave no referido processo de coincidência é que a evolução no tempo em sons coincidentes deve seguir uma correspondência linear, assumindo que as bases de tempo em ambos os lados estão estáveis. Isto é quase sempre verdadeiro, a menos que um dos sons tenha sido deformado intencionalmente de forma não linear ou sujeito a equipamento de reprodução defeituoso tal como um toca-fitas com um problema de velocidade de oscilação. Assim, os pares de marcos

corretos (marco_n , marco_n^*) na lista de dispersão de um dado sound_ID deve ter uma correspondência linear da forma

$$\text{marco}_n^* = m \cdot \text{marco}_n + \text{desvio},$$

onde m é a inclinação, que deve ser perto de um; marco_n é o ponto de tempo dentro da amostra exógena; marco_n^* é o ponto de tempo correspondente dentro da gravação de som indexada pelo sound_ID ; e o desvio é o desvio em tempo na gravação de som correspondendo ao começo da amostra de som exógena. Pares de marcos que podem ser encaixados com a equação acima para valores particulares de m e desvio são relatados como sendo relacionados linearmente. Obviamente, o conceito de ser relacionado linearmente é válido somente para mais de um par de marcos correspondentes. Observa-se que tal relação linear identifica o arquivo de som correto com alta probabilidade, enquanto exclui os pares de marcos afastados que não possuem qualquer significado. Enquanto é possível que dois sinais distintos contenham um número de impressões digitais idênticas, é muito improvável que tais impressões digitais tenham as mesmas evoluções no tempo relativas. A exigência por correspondências lineares é uma característica chave da presente invenção, e provê um reconhecimento melhor significativamente do que técnicas que simplesmente contam o número total de características em comum ou medem a similaridade entre as características. De fato, devido a esse aspecto da invenção, os sons podem ser reconhecidos, mesmo se menos de 1% das impressões digitais da gravação original aparecer na amostra de som exógena, isto é, se a amostra de som for muito curta ou se for significativamente distorcida.

O problema de determinar se há uma coincidência para a amostra exógena é, assim, reduzido ao equivalente de encontrar uma linha diagonal com inclinação perto de um dentro de uma representação de dispersão dos pontos de marcos de uma dada lista de dispersão. Duas representações de dispersão de amostra são ilustradas nas figuras 10A e 10B, com marcos de arquivo de som no eixo horizontal e marcos de amostra de som exógena no eixo vertical. Na

figura 10A, uma linha diagonal de inclinação aproximadamente igual a um é identificada, indicando que a música na verdade coincide com a amostra, isto é, que o arquivo de som constitui um arquivo vencedor. A interceptação no eixo horizontal indica o desvio naquele arquivo de áudio em que a amostra começa.

- 5 Nenhuma linha diagonal estatisticamente significativa é encontrada no diagrama de dispersão da figura 10B, indicando que aquele arquivo de som não é uma coincidência para aquela amostra exógena.

Existem muitas formas de se encontrar uma linha diagonal em um diagrama de dispersão, todas das quais estão dentro do escopo da presente invenção. Deve-se entender que a frase “localizar uma linha diagonal” refere-se a todos os métodos que são equivalentes à localização de uma linha diagonal sem produzir explicitamente uma linha diagonal. Um método preferido inicia subtraindo-se $m \cdot \text{marco}_n$ de ambos os lados da equação acima para resultar em

$$(\text{marco}_n^* - m \cdot \text{marco}_n) = \text{desvio}.$$

Assumindo-se que m é aproximadamente igual a um, a saber,

15 assumindo-se que não haja prolongamento de tempo, chega-se em

$$(\text{marco}_n^* - \text{marco}_n) = \text{desvio}.$$

O problema de encontrar diagonal é então reduzido ao encontrar múltiplos pares de marcos para um dado `sound_ID` que agrupa perto do mesmo valor de desvio. Isso pode ser realizado facilmente pela subtração de um marco a partir do outro e coleta de um histograma dos valores de desvio resultantes.

20 O histograma pode ser preparado ordenando-se os valores de desvio resultantes usando um algoritmo de ordenação rápido ou mediante criação de entradas de compartimento “bin” com contadores e inserção numa árvore B. Compartimento de desvio vencedor no histograma contém o número mais alto de pontos. Este compartimento é referido aqui como o pico do histograma. Posto que o desvio

25 deve ser positivo se o sinal de som exógeno estiver totalmente contido dentro do arquivo de som de biblioteca correto, os pares de marcos que resultam em um desvio negativo podem ser excluídos. De forma similar, os desvios além

do terminal do arquivo também podem ser excluídos. O número de pontos no compartimento de desvio vencedor do histograma é notado para cada sound_ID de qualificação. Esse número se torna a marcação para cada gravação de som. A gravação de som na lista candidata com a marcação mais alta é escolhida
5 como a vencedora. O sound_ID vencedor é reportado para um usuário como descrito abaixo para sinalizar o sucesso da identificação. Para impedir a falsa identificação, uma marcação limiar mínima pode ser utilizada para se obter o sucesso do processo de identificação. Se nenhum som da biblioteca tiver uma marcação que exceda o limiar, então não há reconhecimento, e o usuário é então
10 informado.

Se o sinal de som exógeno contiver múltiplos sons, então cada som individual pode ser reconhecido. Nesse caso, múltiplos vencedores estão localizados na varredura de alinhamento. Não é necessário saber que o sinal de som contém múltiplos vencedores, visto que a varredura de alinhamento
15 localizará mais do que um sound_ID com uma marcação que é muito maior do que as marcações restantes. O método de impressão digital usado exhibe, de preferência, boa superposição linear, tal que as impressões digitais individuais possam, enfim, ser extraídas. Por exemplo, um método de impressão digital de espectrograma exhibe superposição linear.

20 Se a amostra de som tiver sido submetida a prolongamento de tempo, então a inclinação não é identicamente igual a um. O resultado de se assumir uma unidade de inclinação numa amostra prolongada no tempo (ao se assumir que as impressões digitais são invariáveis ao prolongamento do tempo) é que os valores de desvio computados não são iguais. Um jeito de se abordar
25 isto e acomodar o prolongamento de tempo moderado é aumentar o tamanho dos “bins” de desvio, isto é, considerar uma faixa de desvios como sendo igual. Em geral, se os pontos não caírem numa linha reta, então os valores de desvio computados são significativamente diferentes, e um leve aumento no tamanho dos “bins” de desvio não resulta num úmero significativo de falso positivos.

Outras estratégias de encontrar linha são possíveis. Por exemplo, uma transformada Radon ou Hough, descrita em T. Risse, “Hough Transform for Line Recognition”, *Computer Vision and Image Processing*, 46, 327-345, 1989, que são bem conhecidas nas técnicas de pesquisa de visão e gráficos de máquina, pode ser utilizada. Na transformada Hough, cada ponto do diagrama de dispersão se projeta numa linha no espaço (inclinação, desvio). O conjunto de pontos no diagrama de dispersão é, desta forma, projetado no espaço duplo de linhas na transformada Hough. Picos na transformada Hough correspondem a interseções das linhas de parâmetro. O pico global de uma transformada como esta de um dado diagrama de dispersão indica o mais alto número de linhas de interseção na transformada Hough, e assim o mais alto número de pontos colineares. Para se permitir uma variação de velocidade de 5%, por exemplo, a construção da transformada Hough pode ser limitada à região onde parâmetro de inclinação varia entre 0,95 e 1,05, economizando, assim, algum esforço de computação.

Busca Hierárquica

Além da etapa de limitação que elimina candidatas com listas de dispersão muito pequenas, aperfeiçoamentos adicionais na eficiência podem ser realizados. Em um aperfeiçoamento como este, o índice de base de dados é segmentado em pelo menos duas partes de acordo com a probabilidade de ocorrência, e apenas os arquivos de som com a mais elevada probabilidade de coincidência da amostra são inicialmente buscados. A divisão pode ocorrer em vários estágios do processo. Por exemplo, lista de índice principal (figura 8C) pode ser segmentada em duas ou mais partes, tal que as etapas 16 e 20 sejam primeiramente realizadas em um dos segmentos. Em outras palavras, arquivos correspondentes a impressões digitais coincidentes são recuperados a partir de apenas uma fração do índice de base de dados, e uma lista de dispersão é gerada a partir desta fração. Se um arquivo de som vencedor não for localizado, então o processo é repetido no restante do índice de base de dados. Em uma outra

implementação, todos os arquivos são recuperados a partir do índice de base de dados, mas a varredura em linha diagonal é feita separadamente nos diferentes segmentos.

Usando essa técnica, a varredura em linha diagonal, uma parte intensiva em termos de computação do método, é realizada primeiro em um pequeno subconjunto dos arquivos de som no índice de base de dados. Visto que a varredura em linha diagonal apresenta um componente de tempo que é aproximadamente linear com relação ao número de arquivos de som sendo varridos, realização de uma busca hierárquica como tal é altamente vantajosa. Por exemplo, assuma que o índice de base de dados de som contém impressões digitais representando 1.000.000 arquivos de som, mas que somente cerca de 1000 arquivos coincidem com as consultas de amostra com alta frequência, por exemplo, 95% das consultas são para 1000 arquivos, enquanto apenas 5% das consultas são para 999.000 arquivos restantes. Assumindo-se uma dependência linear do custo computacional no número de arquivos, o custo é proporcional a 1000, 95% do tempo, e proporcional a 999.000, apenas 5% do tempo. O custo médio é, por conseguinte, proporcional a cerca de 50.900. Uma busca hierárquica dessa forma resulta aproximadamente num fator de 20 economias na carga de computação. Obviamente, o índice de base de dados pode ser segmentado em mais de 2 níveis de hierarquia, por exemplo, um grupo de novos lançamentos, um grupo de músicas recém lançadas, e um grupo de músicas mais antigas e menos populares.

Conforme descrito acima, a busca é primeiro executada em um primeiro subconjunto de arquivos de som, os arquivos de maior probabilidade e, então, somente se a primeira busca falhar, ela é executada em um segundo subconjunto contendo os arquivos remanescentes. Falha da varredura em linha diagonal ocorre se o número de pontos em cada compartimento de desvio não atingir um valor limiar predeterminado. Alternativamente, ambas buscas podem ser realizadas em paralelo (simultaneamente). Se o arquivo de som correto for

localizado em uma busca do primeiro subconjunto, então um sinal é enviado para encerrar a busca do segundo subconjunto. Se o arquivo de som correto não for localizado na primeira busca, então a segunda busca continua até que um arquivo vencedor seja localizado. Essas duas implementações diferentes
5 envolvem compensações de esforço e tempo de computação. Aquela primeira implementação é mais eficiente em termos de computação, mas introduz uma latência leve se a primeira busca falhar, enquanto dita segunda implementação desperdiça esforço de computação se o arquivo vencedor estiver no primeiro subconjunto mas minimiza a latência se não estiver.

10 O objetivo de segmentar a lista é o de estimar a probabilidade de um arquivo de som ser o alvo de uma consulta e o de limitar a busca a tais arquivos mais prováveis de coincidirem com a amostra da consulta. Existem várias formas possíveis de se designar probabilidades e de ordenar as músicas na base de dados, todas das quais se encontram dentro do escopo da presente
15 invenção. De preferência, probabilidades são designadas com base na recência ou na frequência de serem identificadas como o arquivo de som vencedor. A recência é uma medida útil, particularmente para músicas populares, uma vez que os interesses musicais mudam muito rapidamente com o tempo, à medida que novas músicas são lançadas. Após as marcações de probabilidades serem
20 computadas, classificações são designadas aos arquivos, e a lista auto ordena pela classificação. A lista ordenada é então segmentada em dois ou mais subconjuntos para busca. Menor subconjunto pode conter um número de arquivos predeterminado. Por exemplo, se a classificação localizar um arquivo dentro do topo, digamos, os 1000 primeiros arquivos, então o arquivo é colocado na
25 menor busca e mais rápida. Alternativamente, os pontos de corte para os dois subconjuntos podem ser ajustados dinamicamente. Por exemplo, todos arquivos com uma marcação excedendo um valor limiar particular podem ser colocados dentro do primeiro subconjunto, e desse modo o número de arquivos em cada subconjunto muda de forma contínua.

Um jeito particular de computar a probabilidade é incrementar uma marcação do arquivo de som por um, cada vez que o mesmo é identificado como uma coincidência para a amostra da consulta. Para levar a recência em consideração, todas daquelas marcações são reduzidas periodicamente de modo descendente, de modo que consultas mais recentes tenham um efeito mais forte na classificação do que consultas mais antigas. Por exemplo, todas marcações podem ser ajustadas descendente por um fator de multiplicação constante no momento de cada consulta, resultando em uma diminuição exponencial da marcação se não for atualizada. Dependendo do número de arquivos na base de dados, que pode ser facilmente de um milhão, esse método pode requerer a atualização de um grande número de marcações em cada consulta, tornando a mesma potencialmente indesejável. Alternativamente, as marcações podem ser ajustadas de forma descendente em intervalos relativamente infrequentes, tais como uma vez por dia. O resultado da ordenação do ajuste menos frequente é efetivamente similar, mas não tão idêntico, à ordenação resultante do ajuste com cada consulta. Contudo, a carga de computação para atualizar as classificações é muito menor.

Uma leve variação desse ajuste de recência, que preserva mais exatamente a marcação de recência, é adicionar uma atualização de marcação exponencialmente crescente a^t ao arquivo de som vencedor mediante consulta “query”, onde t é a quantidade de tempo que passou desde a última atualização global. Todas as marcações são então ajustadas para baixo pela divisão por a^T em cada atualização global, onde T é o tempo total que passou desde a última atualização global. Nesta variação, a é um fator de recência que é superior a um.

Em adição à classificação descrita acima, algum conhecimento *a priori* pode ser introduzido para auxiliar a semear a listagem. Por exemplo, novos lançamentos devem ter números mais altos de consultas do que músicas antigas. Assim, novos lançamentos podem ser colocados automaticamente no primeiro subconjunto contendo sons com uma maior probabilidade de coincidir

consultas. Isso pode ser executado independentemente do algoritmo de auto-classificação descrito acima. Se a característica de autoclassificação também for utilizada, novos lançamentos podem ser designados com classificações iniciais que colocam as mesmas em algum lugar dentro do primeiro subconjunto. Os

5 novos lançamentos podem ser semeados no topo da lista, no fundo da lista de músicas de maior probabilidade, ou em algum lugar entre os mesmos. Para os fins da busca, a localização inicial não importa, já que a classificação converge com o tempo para refletir o verdadeiro nível de interesse.

Em uma modalidade alternativa, a busca é realizada na ordem

10 das classificações de recência e é encerrada quando uma marcação `sound_ID` excede um valor limiar predeterminado. Isso é equivalente ao método acima, no qual cada segmento contém um `sound_ID` apenas. Experimentos ilustram que a marcação de um som vencedor é muito mais alta do que marcações de todos os outros arquivos de som, e, dessa forma, um limiar adequado pode ser

15 escolhido com a experimentação mínima. Uma forma de se implementar essa modalidade é a de classificar todos os `sound_IDs` no índice de base de dados de acordo com a recência, com um desempate arbitrário no caso de marcações idênticas. Uma vez que cada classificação de recência é singular, existe um mapeamento de um para um entre a marcação de recência e aquele `sound_ID`.

20 A classificação pode então ser usada, em vez do `sound_ID`, quando ordenando por `sound_ID` para formar a lista de `sound_IDs` candidatas e listas de dispersão associadas (figura 9C). Os números das classificações podem ser limitados ao índice quando a lista de índice de tríplex (impressão digital, marco, `sound_ID`) é gerada e antes da lista de índice ser ordenada numa lista de índice principal.

25 A classificação então toma o lugar daquele `sound_ID`. Alternativamente, uma função de busca e de substituição pode ser utilizada para substituir o `sound_ID` com a classificação. À medida que ditas classificações são atualizadas, novas classificações são mapeadas nas classificações antigas, desde que a integridade de mapeamento seja mantida.

Alternativamente, as classificações podem ser limitadas depois no processo. Uma vez que as listas de dispersão são criadas, uma classificação pode ser associada com cada `sound_ID`. Os conjuntos são então ordenados por classificação. Nessa implementação, só os apontadores das listas de dispersão
5 precisam ser modificados; o agrupamento nas listas de dispersão não precisa ser repetido. A vantagem de uniões posteriores é que todo o índice de base de dados não precisa ser recriado toda vez que as classificações são atualizadas.

Note-se que a classificação na lista de popularidade pode por si só ser de interesse como um objeto de valor econômico. Isto é, tal classificação
10 reflete o desejo dos consumidores de obter uma identificação de uma amostra de música desconhecida. Em muitos casos, a consulta é criada por um desejo de se comprar uma gravação da música. De fato, se a informação demográfica sobre o usuário é conhecida, então os esquemas de classificação alternativos podem ser implementados para cada grupo demográfico desejado. Um grupo
15 demográfico de usuário pode ser obtido daquela informação de perfil solicitada quando o usuário se inscreve no serviço de reconhecimento. Isso também pode ser determinado dinamicamente por técnicas de filtragem colaborativa padrão.

Em um sistema em tempo real, o som é fornecido ao sistema de reconhecimento de maneira incremental ao longo do tempo, possibilitando o
20 reconhecimento canalizado. Neste caso, é possível processar o dado que entra em segmentos e atualizar o conjunto de índices de amostra de modo incremental. Após cada período de atualização, o conjunto de índices recém aumentado é utilizado para recuperar gravações de som da biblioteca candidata utilizando as etapas de busca e varredura acima. Tal índice de base de dados é buscado
25 para impressões digitais que coincidem com impressões digitais de amostras recém obtidas, e as novas trípticas (marco_k , marco^*_j , `sound_IDj`) são geradas. Novos pares são acrescentados às listas de dispersão, e os histogramas são aumentados. A vantagem de dita abordagem é que se dados suficientes forem coletados para identificar gravação de som de modo não ambíguo, por exemplo,

se o número de pontos num compartimento de desvio de um dos arquivos de som exceder um limiar alto ou exceder a próxima marcação de arquivo de som mais alta, então a aquisição de dados pode ser encerrada e resultado anunciado.

Uma vez que o som correto tem sido identificado, o resultado
5 é reportado para um usuário ou o sistema por qualquer método adequado. Por exemplo, o resultado pode ser reportado por um impresso de computador, e-mail, página de resultado de busca na rede, mensagem de texto SMS (serviço de envio de mensagens curtas) para um telefone móvel, anotação de voz gerada por computador através de um telefone, ou postagem do resultado em um site
10 da rede ou conta da Internet que o usuário possa acessar depois. Os resultados reportados podem incluir informação de identificação do som tal como o nome do artista de uma música; o nome do compositor, e atributos de gravação (por exemplo, intérpretes, regentes, “venue”) de uma peça clássica; a companhia e o produto de um anúncio; ou quaisquer outros identificadores adequados. Ademais,
15 informação biográfica, informação sobre concertos na cercania, e outra informação de interesse para fãs podem ser providas; hyperlinks para tais dados podem ser providos. Os resultados relatados também podem incluir a marcação absoluta do arquivo de som ou sua marcação em comparação com o próximo mais elevado arquivo marcado.

20 Um resultado útil do método de reconhecimento é que isto não confunde duas rendições diferentes do mesmo som. Por exemplo, diferentes execuções da mesma peça de música clássica não são consideradas as mesmas, mesmo se um ser humano não conseguir detectar uma diferença entre as duas. Isso é pelo fato de ser altamente improvável que os pares de marco/impressão
25 digital e evolução delas no tempo coincidam exatamente para duas execuções diferentes. Numa modalidade atual, os pares de marco/impressão digital devem estar em cerca de 10 ms um do outro para que uma correspondência linear seja identificada. Como resultado, reconhecimento automático efetuado pela invenção torna possível que a execução/trilha sonora adequada e o artista/legenda sejam

creditados em todos os casos.

Implementação Exemplar

Uma implementação preferencial da invenção, reconhecimento de áudio de janela deslizante contínua, está descrita abaixo. Um microfone ou
5 outra fonte de som é continuamente amostrado num armazenamento temporário “buffer” para obter um registro dos N segundos anteriores do som. O conteúdo do armazenamento temporário do som é periodicamente analisado para garantir a identidade do conteúdo do som. O armazenamento temporário do som pode ter um tamanho fixo ou pode crescer em termo de tamanho à medida que o som
10 é amostrado, referido aqui como segmentos sequencialmente crescentes daquela amostra de áudio. Um relatório é feito para indicar a presença de gravações de som identificadas. Por exemplo, um arquivo de registro “log” pode ser coletado, ou uma exibição pode ser mostrada em um dispositivo indicando informações sobre a música, tal como o título, o artista, a arte da capa do álbum, letras, ou
15 informações de compra. Para evitar redundância, um relatório pode ser criado apenas quando a identidade do som reconhecido mudar, por exemplo, depois de uma mudança do programa num juke-box. Um dispositivo desse tipo pode ser usado para criar uma lista de músicas tocadas a partir de qualquer corrente de som (rádio, rádio “streaming” pela Internet, microfone escondido, chamada telefônica, etc.). Além da identidade musical, informação tal como o tempo de reconhecimento pode ser registrado. Se dita informação de localização estiver disponível (por exemplo, via GPS), tal informação também pode ser registrada.

Para realizar a identificação, cada “buffer” pode ser identificado *de novo*. Alternativamente, parâmetros de som podem ser extraídos, por exemplo,
25 em impressões digitais ou outras formas extraídas de característica intermediária, e armazenados em um segundo “buffer”. Novas impressões digitais podem ser adicionadas à frente do segundo “buffer”, com as impressões digitais antigas sendo descartadas do terminal do “buffer”. A vantagem de um esquema do tipo “buffer” de rolagem como este é que a mesma análise não precisa ser efetuada

de modo redundante em segmentos de sobreposição antigos de amostras de som, economizando desta forma esforço de computação. O processo de identificação é periodicamente feito sobre o conteúdo do “buffer” de rolagem de impressão digital. No caso de um dispositivo portátil pequeno, a análise de impressão digital

5 pode ser efetuada no dispositivo e os resultados transmitidos para um servidor de reconhecimento que usa um canal de dados de largura de banda relativamente baixa, já que a corrente de impressão digital não é muito intensa em termos de dados. Aquele “buffer” de rolagem de impressão digital pode ser mantido no dispositivo portátil e transferido para o servidor de reconhecimento a cada vez,

10 ou pode ser mantido no servidor de reconhecimento, caso no qual uma sessão de reconhecimento contínua é armazenada temporariamente no servidor.

Num sistema de reconhecimento de “buffer” de rolagem desse tipo, novas gravações de sons podem ser reconhecidas tão logo a informação suficiente estiver disponível para reconhecimento. Informação suficiente pode

15 ocupar menos do que o comprimento do “buffer”. Por exemplo, se uma música distintiva puder ser reconhecida unicamente após um segundo de reprodução, e o sistema tiver uma periodicidade de reconhecimento de um segundo, então a música pode ser reconhecida imediatamente, embora o “buffer” possa ter uma duração de 15-30 segundos. Por outro lado, se uma música menos distintiva exigir

20 mais segundos de amostra para reconhecê-la, o sistema deve aguardar um período mais longo antes de declarar, enfim, a identidade da música. Nesse esquema de reconhecimento de janela deslizante, músicas são reconhecidas assim que elas puderem ser identificadas.

É importante notar que, enquanto a presente invenção tem sido

25 descrita no contexto de um sistema e um método de reconhecimento totalmente funcional, os versados na técnica irão apreciar que o mecanismo da invenção é capaz de ser distribuído na forma de meio legível por computador de instruções em várias formas, e que a invenção é aplicável igualmente, independentemente do tipo particular da mídia que porta sinal usado para efetuar de fato a distribuição.

Exemplos de tais dispositivos acessíveis por computador incluem memória de computador (RAM ou ROM), discos flexíveis, e CD-ROMs, assim como uma mídia tipo transmissão tal como enlaces de comunicações digitais e analógicas.

REIVINDICAÇÕES

1. Método para reconhecer uma amostra de áudio, caracterizado pelo fato de compreender as etapas de:

(a) para a amostra, gerar pares de marcos/impressões digitais de amostras, em que cada marco ocorre em uma localização particular no tempo dentro da amostra de áudio, a localização sendo computada na dependência do conteúdo da amostra de áudio, e em que cada impressão digital caracteriza uma ou mais características da amostra de áudio na ou perto da localização particular;

(b) para cada um de um ou mais arquivos de áudio, gerar uma lista de pares de correspondência de marco;

(c) identificar correspondências substancialmente lineares entre os respectivos pares de marcos/impressões digitais de amostras e os pares de marcos/impressões digitais de arquivos gerados previamente; e

(d) identificar um arquivo vencedor, tal como aquele que possui um número significativo de correspondências substancialmente lineares.

2. Método de acordo com a reivindicação 1, caracterizado pelo fato de incluir um estágio de geração de impressão digital de arquivo, no qual para cada um dentre um ou mais arquivos de áudio, pares de marcos/impressões digitais de arquivos são gerados, em que cada marco ocorre numa localização particular no tempo dentro do arquivo de áudio, a localização sendo computada na dependência do conteúdo do arquivo de áudio, e em que cada impressão digital caracteriza uma ou mais características do arquivo de áudio na ou perto da localização particular.

3. Método de acordo com a reivindicação 1 ou 2, caracterizado pelo fato de cada impressão digital representar um número de características do áudio em cada localização de marco, ou desviar levemente da localização.

4. Método de acordo com qualquer uma das reivindicações 1 a 3, caracterizado pelo fato de cada impressão digital ser computada por meio de um método de modo que seja invariável a prolongamento de tempo da amostra.

5. Método de acordo com qualquer uma das reivindicações 1 a 4, caracterizado pelo fato de cada impressão digital ser calculada como uma de uma impressão digital de fatia espectral, uma impressão digital de múltiplas fatias, um coeficiente LPC, um coeficiente “cepstral”, e um componente de frequência de picos de espectrograma.

6. Método de acordo com a reivindicação 5, caracterizado pelo fato de que uma impressão digital de fatia espectral é calculada num conjunto de desvios de tempo para um ponto de tempo do marco.

7. Método de acordo com qualquer uma das reivindicações 1 a 6, caracterizado pelo fato de que a posição de cada marco é identificada usando um método de formação de marcos que encontra as localizações distintivas e reproduzíveis dentro da gravação de som.

8. Método de acordo com a reivindicação 7, caracterizado pelo fato de que o método de formação de marcos usa uma norma L_p espectral para calcular a energia instantânea em cada ponto de tempo possível na gravação e seleciona máxima local como os marcos.

9. Método de acordo com a reivindicação 7 ou 8, caracterizado pelo fato de que um ou mais marco(s) é um marco de múltiplas fatias derivado de componentes espectrais através de múltiplos intervalos de tempo em desvios fixos ou variáveis um do outro.

10. Método de acordo com qualquer uma das reivindicações 1 a 9, caracterizado pelo fato de os pares de marcos/impressões digitais de arquivos estarem armazenados numa base de dados, cada arquivo dentro da base de dados sendo indexado por impressões digitais daquele arquivo.

11. Método de acordo com a reivindicação 10, caracterizado pelo fato de que os índices são ordenados de acordo com a impressão digital.

12. Método de acordo com a reivindicação 11, caracterizado pelo fato de uma lista de índice principal ser compilada tendo uma entrada para cada impressão digital única e apontador para uma lista de marcos correspondentes.

13. Método de acordo com qualquer uma das reivindicações 10 a 12, caracterizado pelo fato de cada arquivo ser identificado por um sound_ID, e a base de dados armazenar uma pluralidade de impressões digitais, marcos, tríplexes de sound_ID.

5 14. Método de acordo com qualquer uma das reivindicações 1 a 13, caracterizado pelo fato de que o arquivo, com os pares de correspondência mais significativos estatisticamente e relacionados linearmente, é selecionado como o arquivo vencedor.

10 15. Método de acordo com qualquer uma das reivindicações 1 a 14, caracterizado pelo fato de que uma correspondência linear entre amostra e marcos de arquivo (marco_n , marco_n^*) ocorre quando um par de marco/impressão digital de amostra coincide com um par de marco/impressão digital de arquivo dentro de uma faixa de tolerância.

15 16. Método de acordo com qualquer uma das reivindicações 1 a 15, caracterizado pelo fato de que uma correspondência linear entre um par de marco/impressão digital de amostra e um par de marco/impressão digital de arquivo ocorre quando as respectivas impressões digitais são coincidadas e os respectivos marcos são relacionados por uma relação linear.

20 17. Método de acordo com a reivindicação 16, caracterizado pelo fato de que tais impressões digitais são coincidadas quando elas forem idênticas ou diferirem dentro de uma tolerância previamente determinada.

25 18. Método de acordo com a reivindicação 16 ou reivindicação 17, caracterizado pelo fato de que uma correspondência linear ocorre se o par de amostra/marco de arquivo (marco_n , marco_n^*) dentro da lista está relacionado de acordo com a relação: $\text{marco}_n^* = m \cdot \text{marco}_n + \text{desvio}$.

19. Método de acordo com qualquer uma das reivindicações 1 a 18, caracterizado pelo fato de que a amostra está na forma de ondas acústicas, ondas de rádio, uma corrente PCM de áudio digital, uma corrente comprimida de áudio digital, ou uma difusão de “streaming” pela Internet.

20. Método de acordo com qualquer uma das reivindicações 1 a 19, caracterizado pelo fato de que ditas impressões digitais de amostras estão armazenadas num “buffer” de rolagem.

5 21. Método de acordo com a reivindicação 20, caracterizado pelo fato de que a etapa de identificar pode ser executada periodicamente sobre os conteúdos do “buffer” de rolagem de impressão digital.

22. Método de acordo com a reivindicação 20 ou reivindicação 21, caracterizado pelo fato de a etapa de identificar poder ser executada assim que informação suficiente esteja disponível no “buffer” para reconhecimento.

10 23. Método de acordo com qualquer uma das reivindicações 1 a 22, caracterizado pelo fato de a etapa de identificar ser primeiro realizada num subconjunto de arquivos, e caso nenhum arquivo vencedor seja identificado no primeiro subconjunto, um segundo subconjunto contendo os arquivos restantes é buscado.

15 24. Método de acordo com a reivindicação 23, caracterizado pelo fato de que o primeiro subconjunto contém arquivos que possuem uma maior probabilidade *a priori* ou empírica de serem identificados do que os arquivos que não estão no primeiro subconjunto.

20 25. Método de acordo com a reivindicação 1, caracterizado pelo fato de tal etapa de identificar correspondências lineares compreender localizar uma linha em diagonal dentro de um diagrama de dispersão de tais localizações correspondentes formando-se diferenças entre tais localizações correspondentes e calculando-se um pico de um histograma de tais diferenças.

25 26. Método de acordo com a reivindicação 1, caracterizado pelo fato de a etapa de identificar um arquivo vencedor compreender adicionalmente prover um indicador de um desvio a uma localização no dito arquivo vencedor onde o número significativo de correspondências ocorrem.

27. Método de acordo com a reivindicação 1, caracterizado pelo fato de que as ditas etapas (a) até (d) do método são realizadas em um servidor,

e em que o dito método compreende ainda as etapas de:

em resposta a uma solicitação de um cliente, retransmitir pelo menos uma porção da amostra de áudio para o servidor; e

5 em resposta ao dito servidor, identificar um arquivo vencedor, respondendo ao dito cliente consequentemente.

28. Sistema de computador para realizar um método conforme definido em qualquer uma das reivindicações 1 a 27, caracterizado pelo fato de incluir um terminal de cliente que envia um sumário extraído de característica da amostra de sinal capturada contendo pares de marcos e impressões digitais
10 para um terminal de servidor, o qual realiza o reconhecimento.

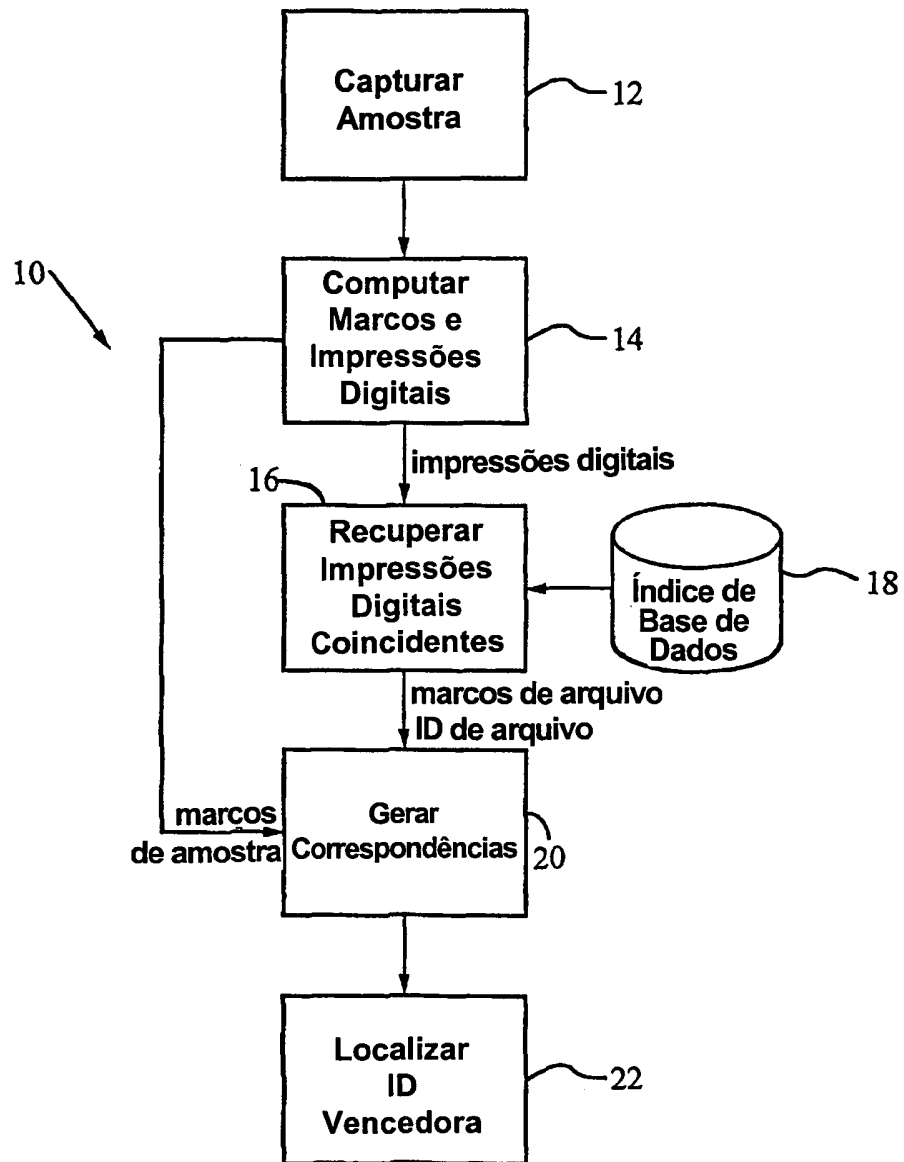
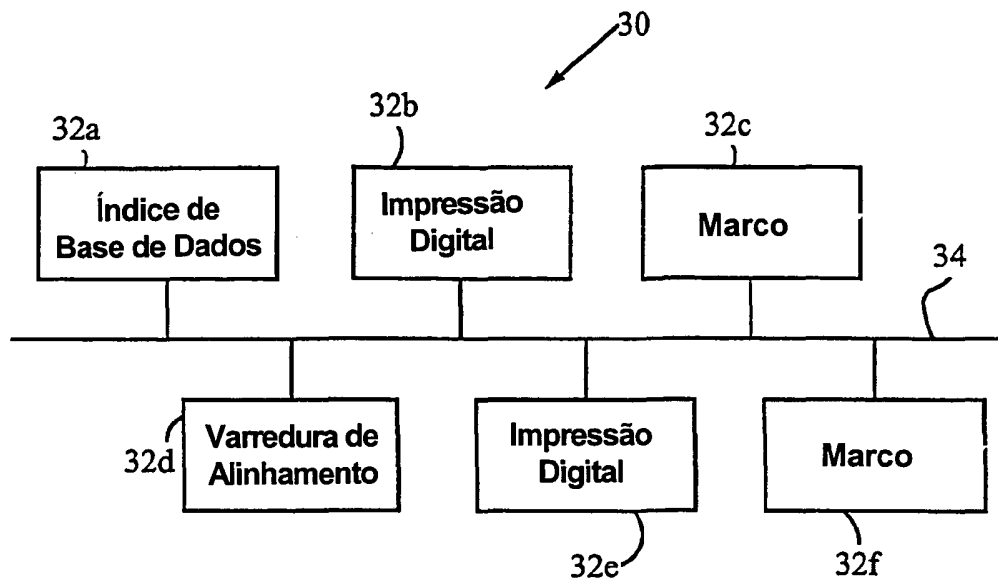
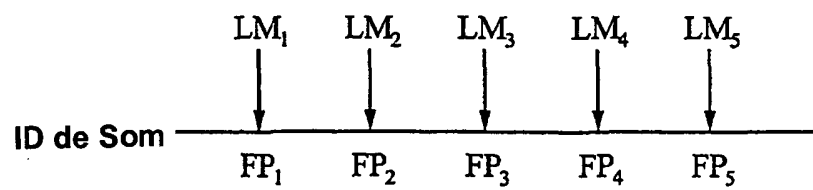


FIG. 1

**FIG. 2****FIG. 4**

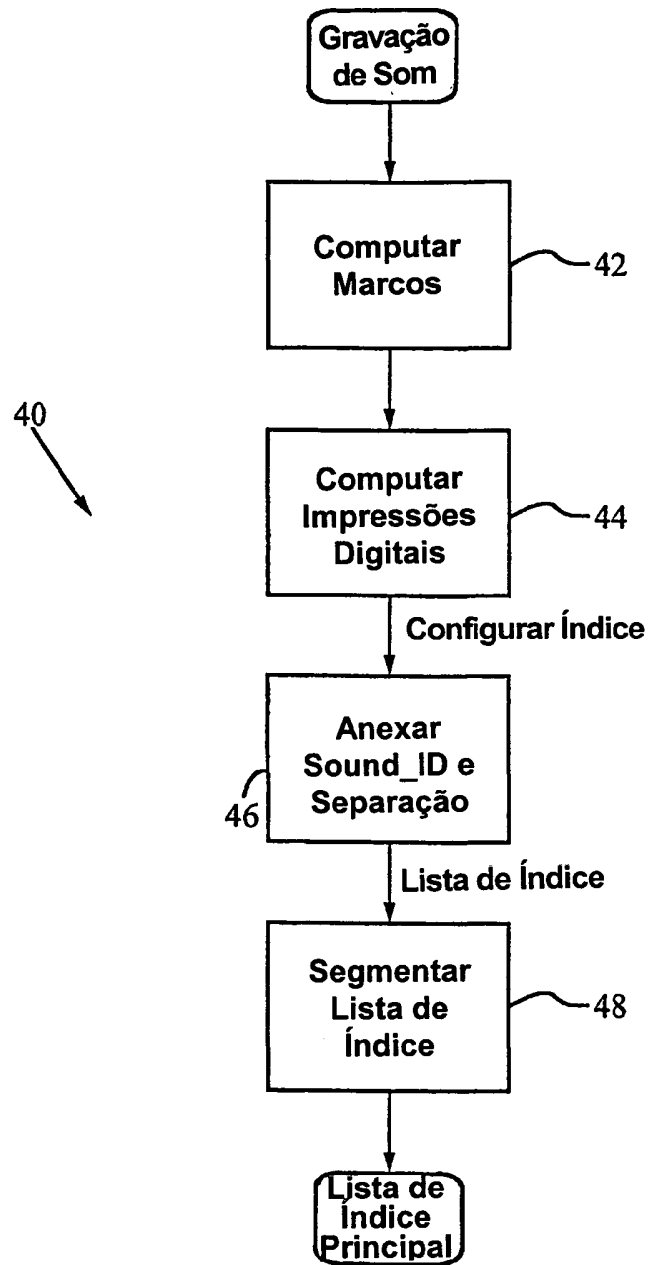


FIG. 3

Marcos na Norma L4

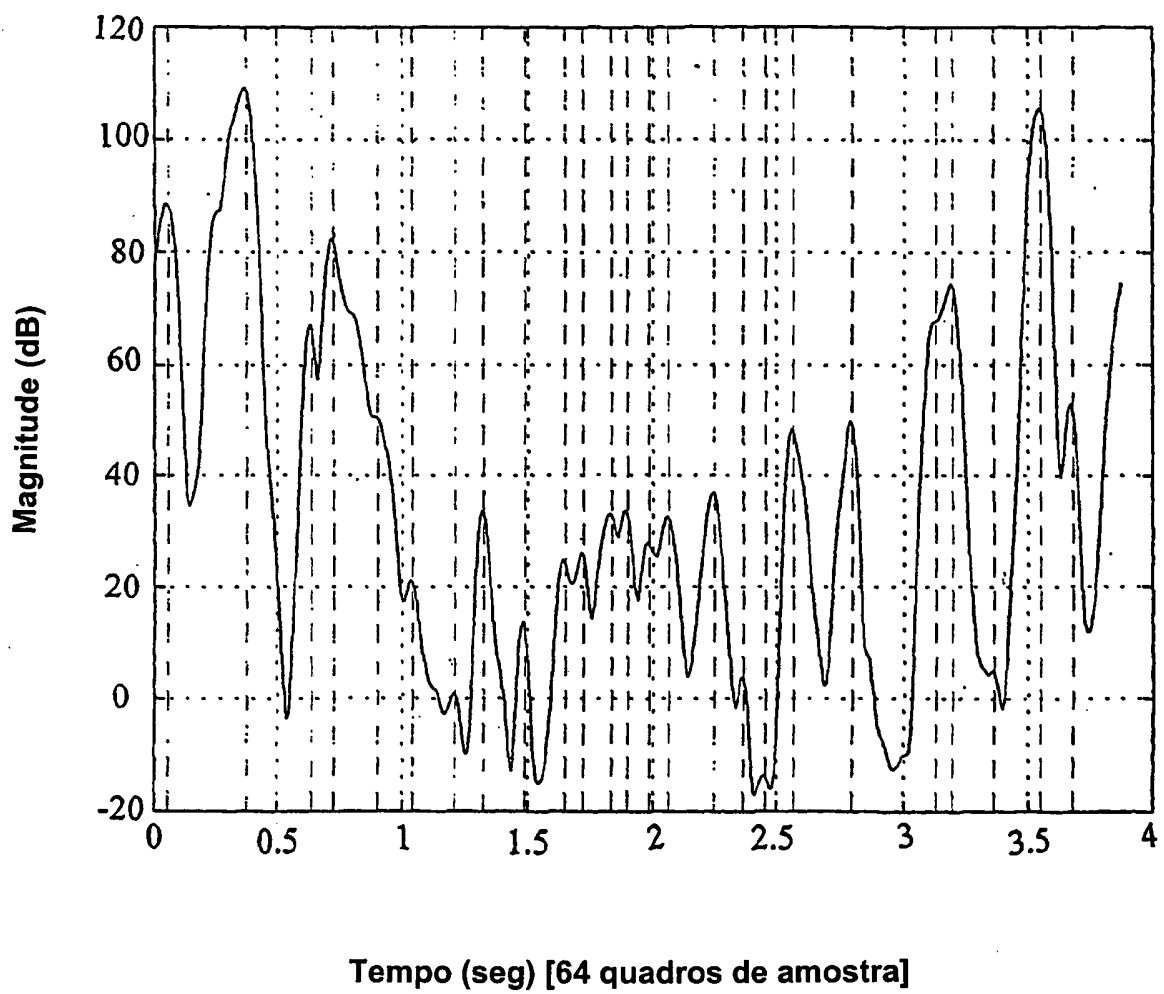


FIG. 5

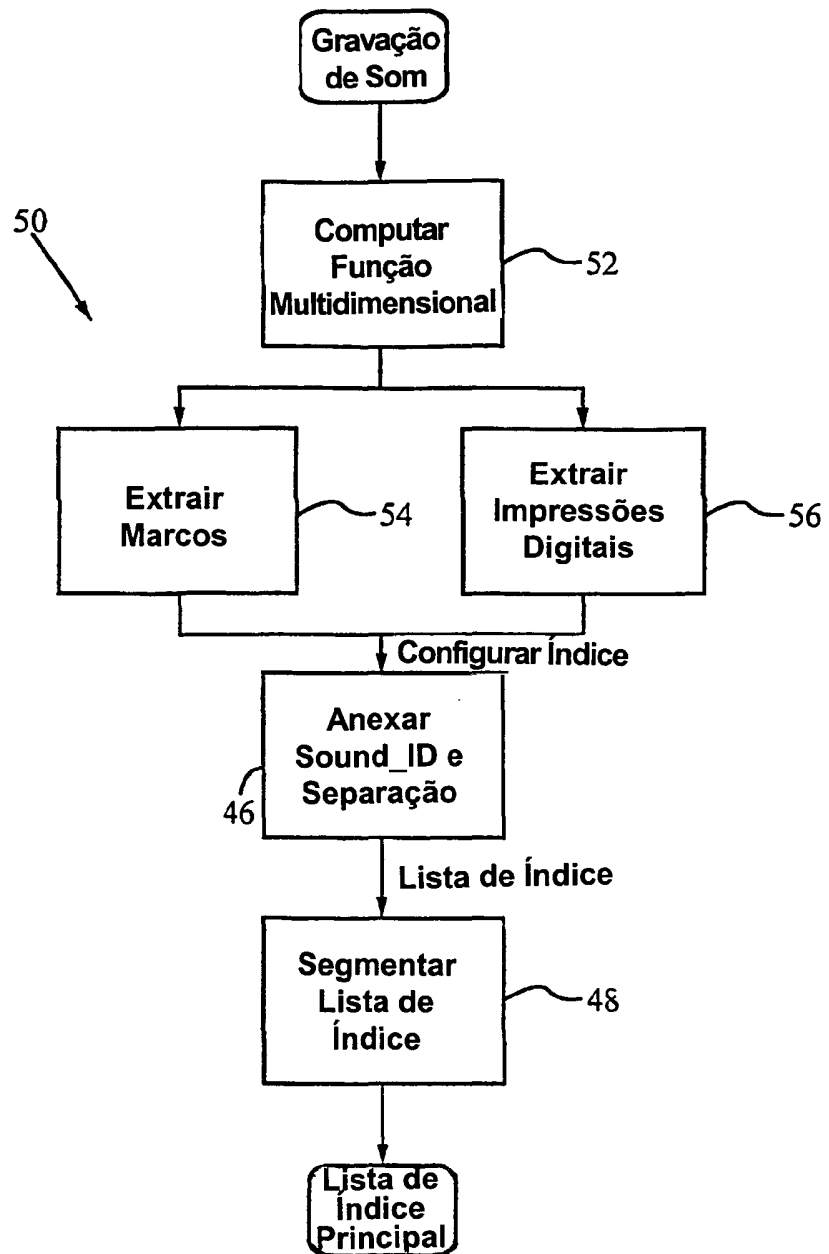
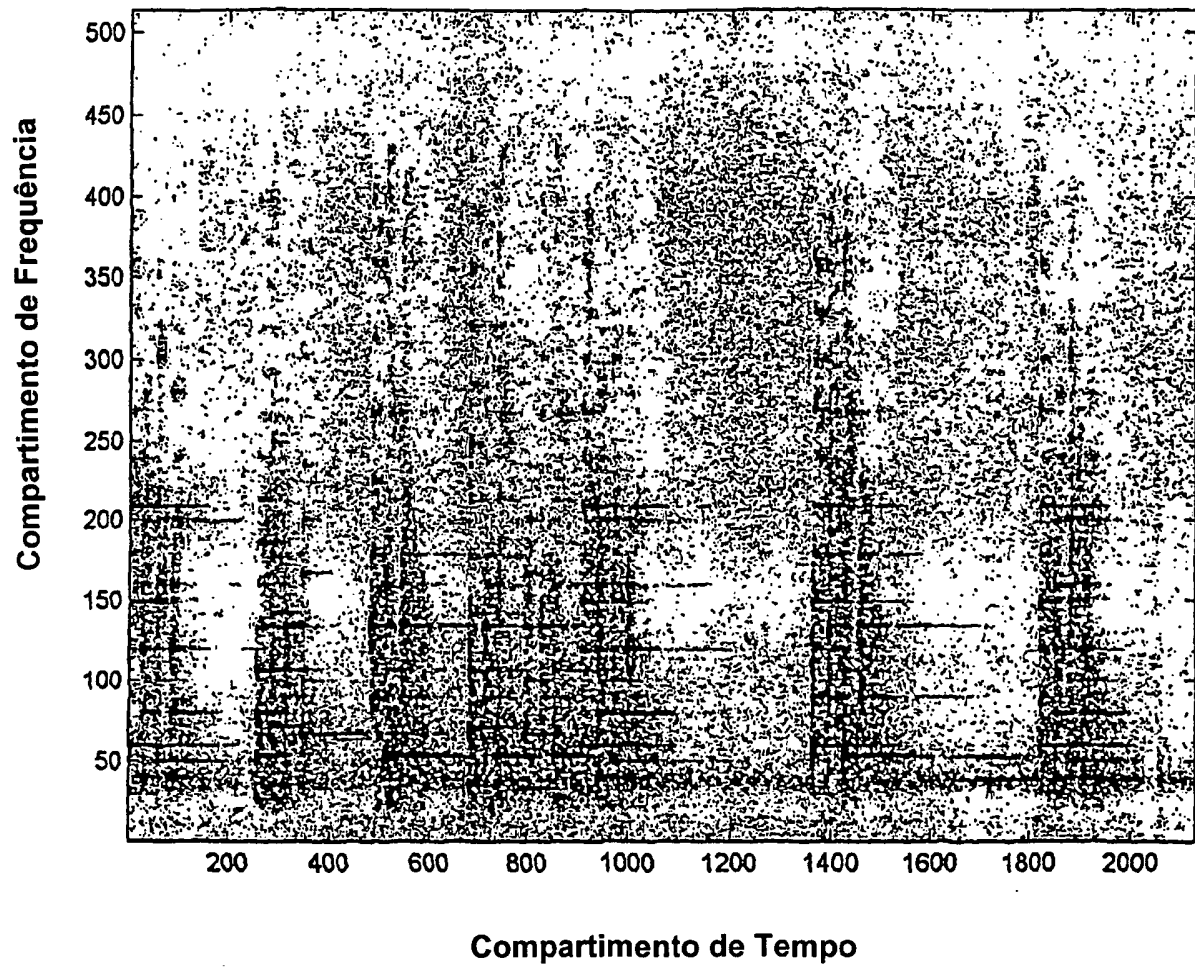
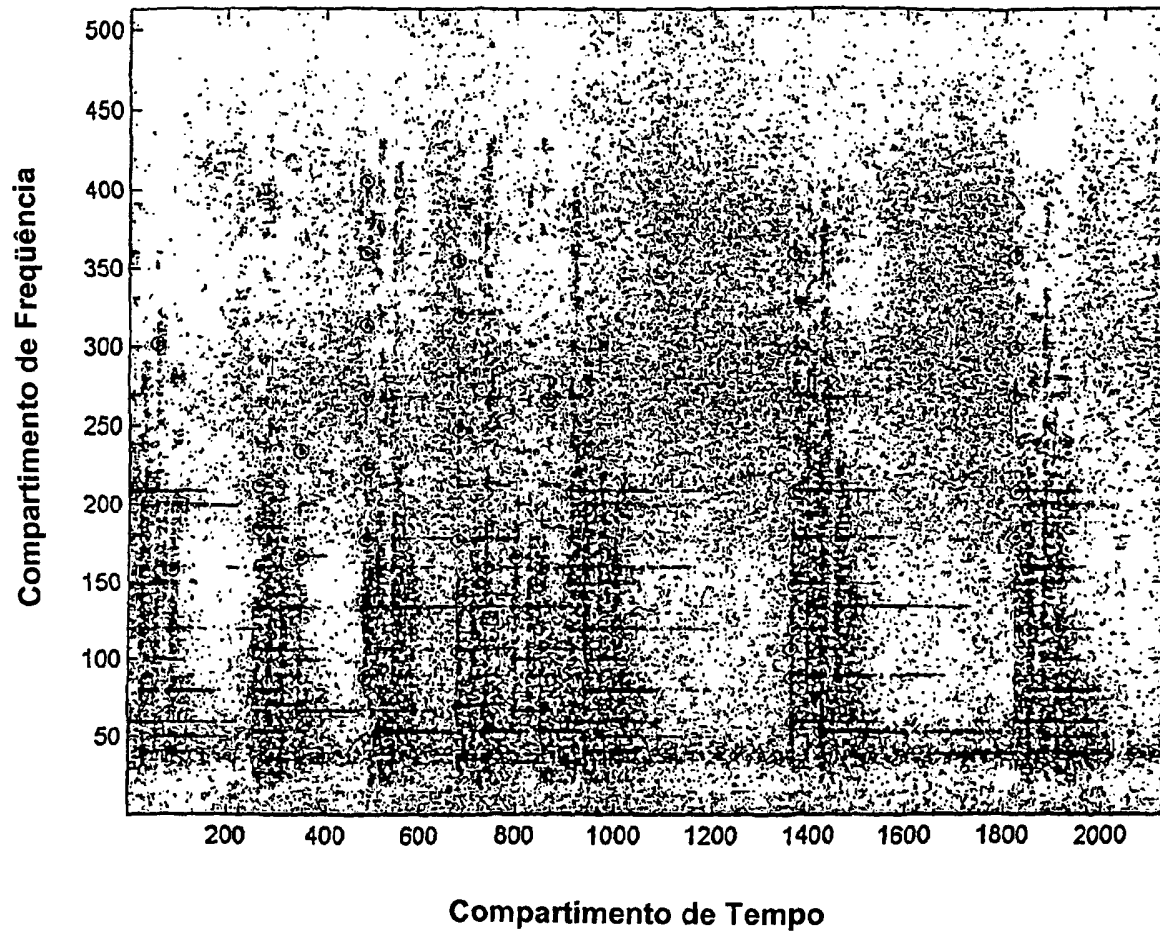
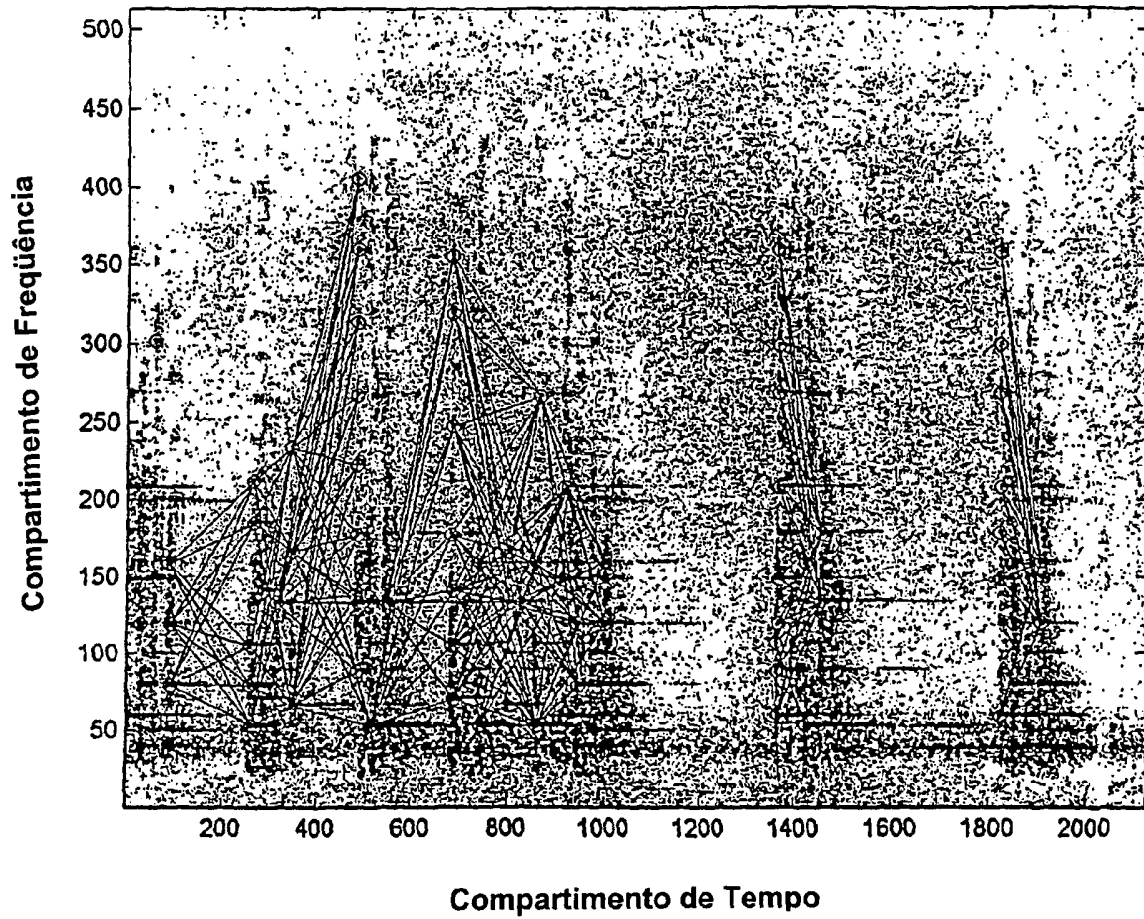
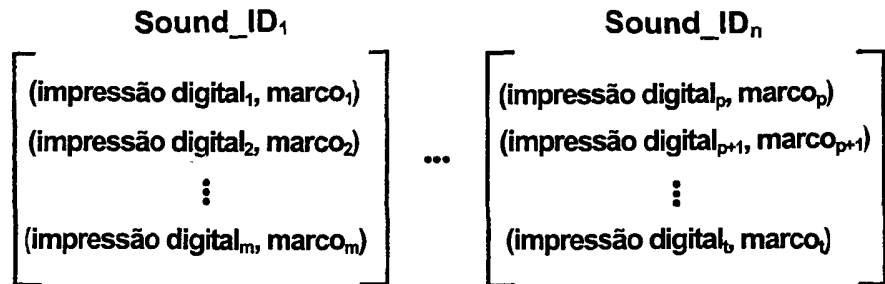
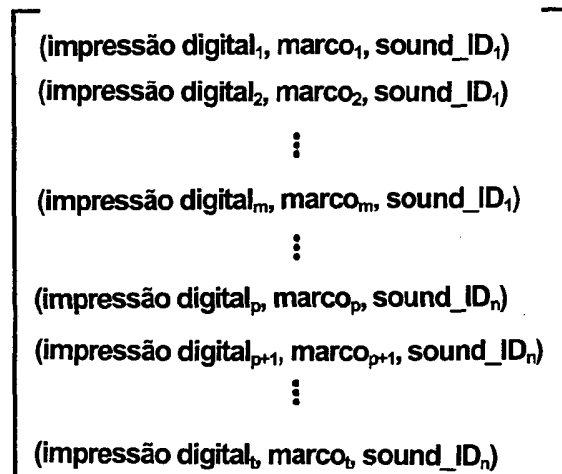


FIG. 6

**FIG. 7A**

***FIG. 7B***

*FIG. 7C*

Configurações de Índice**FIG. 8A**Lista de Índice**FIG. 8B**

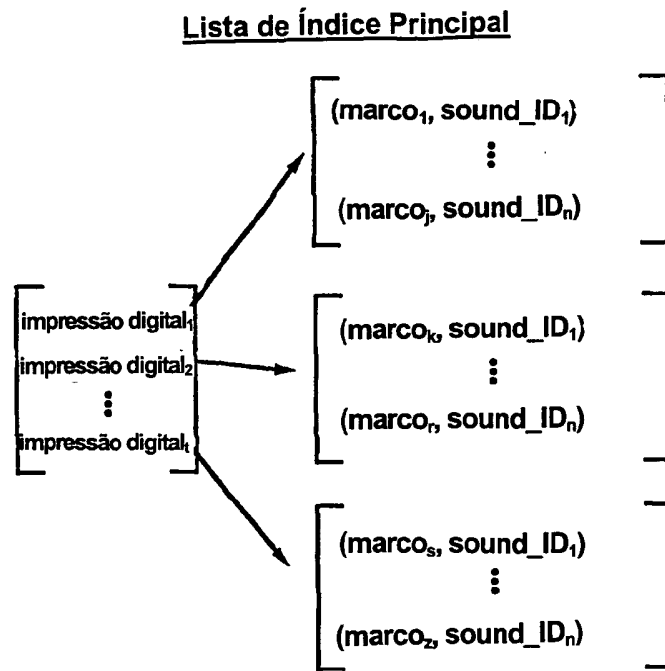


FIG. 8C

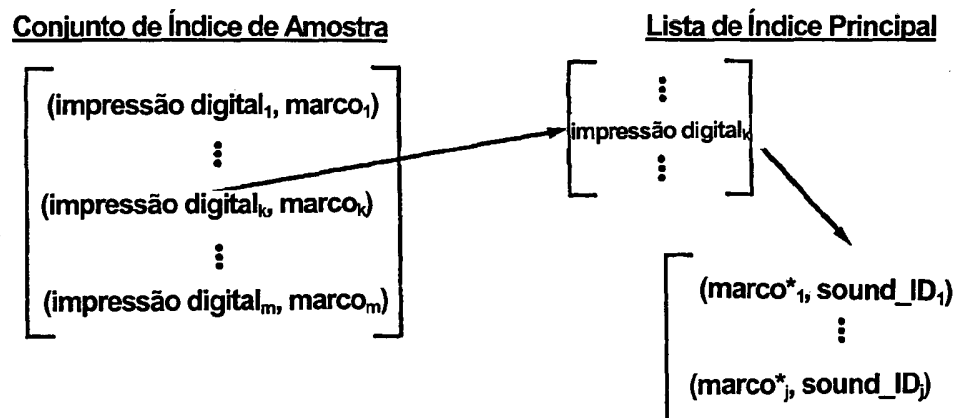
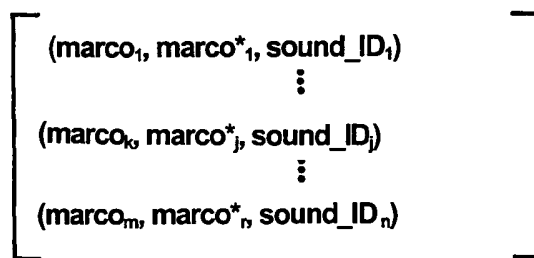
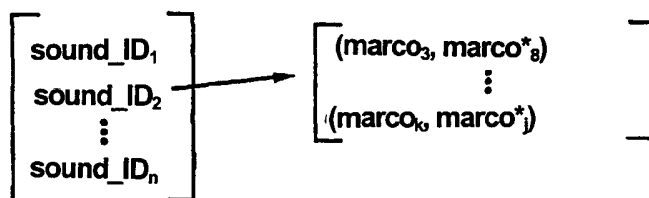


FIG. 9A

Lista Candidata**FIG. 9B**ID de Som CandidatasLista de Espalhamento**FIG. 9C**

Arquivo vencedor localizado: correspondência linear encontrada

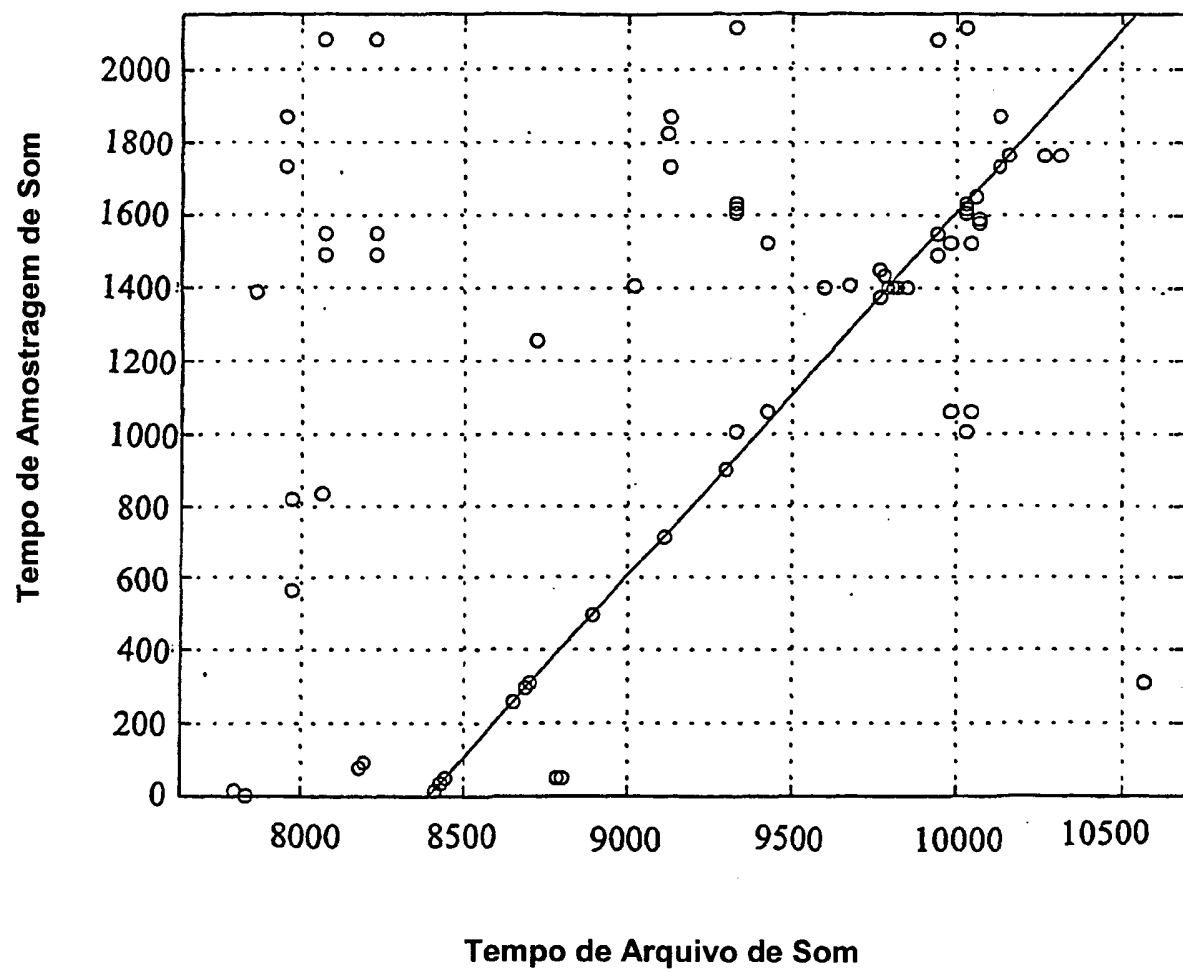


FIG. 10A

Arquivo vencedor não localizado: correspondência linear não encontrado

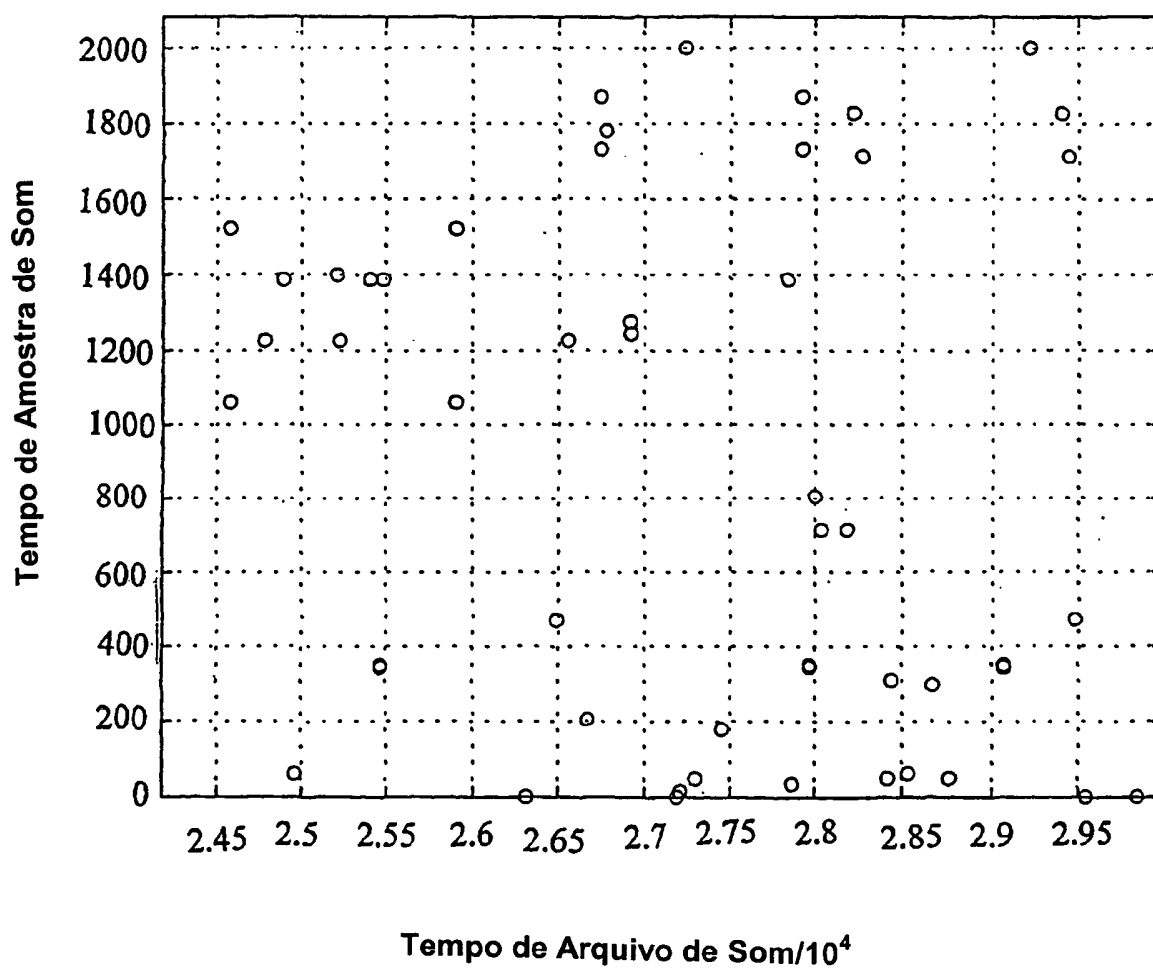


FIG. 10B

RESUMO

“MÉTODO PARA RECONHECER UMA AMOSTRA DE ÁUDIO, E, SISTEMA DE COMPUTADOR PARA REALIZAR O MESMO”

Um método para reconhecer uma amostra de áudio localiza um
5 arquivo de áudio que mais proximamente coincide com a amostra de áudio de
uma base de dados a qual indexa um grande conjunto de gravações originais.
Cada arquivo de áudio indexado é representado no índice de base de dados por
um conjunto de pontos de tempo de marco e impressões digitais associadas. Os
marcos ocorrem em localizações reproduzíveis dentro do arquivo, enquanto as
10 impressões digitais representam características do sinal no ou perto dos pontos
de tempo do marco. Para efetuar reconhecimento, marcos e impressões digitais
são computados para amostra desconhecida e usados para recuperar impressões
digitais coincidentes da base de dados. Para cada arquivo contendo impressões
digitais coincidentes, os marcos são comparados com marcos da amostra, nos
15 quais as mesmas impressões digitais foram computadas. Se um grande número
de marcos correspondentes for relacionado linearmente, isto é, se impressões
digitais equivalentes da amostra e arquivo recuperado tiverem mesma evolução
no tempo, então o arquivo é identificado com a amostra. Este método pode ser
usado para qualquer tipo de som ou música, e é particularmente eficaz para os
20 sinais de áudio sujeitos à distorção linear e não linear tal como ruído de fundo,
artefatos de compressão, ou interrupções de transmissão. A amostra pode ser
identificada num momento proporcional ao logaritmo do número de entradas
na base dedados; dada uma energia computacional suficiente, reconhecimento
pode ser executado quase que em tempo real, à medida que o som está sendo
25 amostrado.