**República Federativa do Brasil**

Ministério do Desenvolvimento, Indústria, Comércio e Serviços

Instituto Nacional da Propriedade Industrial

*BRPI0922035B1*

**(11) PI 0922035-6 B1**

**(22) Data do Depósito:** 12/11/2009

**(45) Data de Concessão:** 24/01/2023

**(54) Título:** MÉTODO DE RECONHECIMENTO DE VOZ E SISTEMA DE RECONHECIMENTO DE VOZ

**(51) Int.Cl.:** G10L 15/16.

**(30) Prioridade Unionista:** 11/11/2009 US 12/616,723.

**(73) Titular(es):** SCTI HOLDINGS, INC..

**(72) Inventor(es):** MARK PINSON; DAVID PINSON, SR.; MARY FLANAGAN; SHAHROKH MAKANVAND.

**(86) Pedido PCT:** PCT US2009064214 de 12/11/2009

**(87) Publicação PCT:** WO 2010/056868 de 20/05/2010

**(85) Data do Início da Fase Nacional:** 12/05/2011

**(57) Resumo:** SISTEMA E MÉTODO PARA CONVERSÃO AUTO MÁTICA DE VOZ PARA TEXTO. Reconhecimento de voz é executado próximo a tempo real e aprimorando pela exploração de eventos e sequência de eventos, empregando técnicas de aprendizado de máquina que incluem classificadores, conjuntos, detectores e cascatas e usando grupos perceptivos. Reconhecimento de voz também é aprimorado usando processamento em série.Um pontuador automático injeta pontuação em sequências de texto reconhecidas.

Relatório Descritivo da Patente de Invenção para **"MÉTODO DE RECONHECIMENTO DE VOZ E SISTEMA DE RECONHECIMENTO DE VOZ".**

Referência Cruzada Para Pedidos Relacionados

[001] Este pedido de patente reivindica o benefício do pedido de patente Norte Americano de número serial 12/616.723, *System and Method for Automatic Speech to Text Conversion*, depositado em 11 de novembro de 2009, e do pedido de patente Norte Americano de número serial 61/113.910, *Automated Speech Processors and Automated Punctuator*, depositado em 12 de novembro de 2008, cujos conteúdos são integralmente incorporados neste documento através desta referência aos mesmos.

Antecedentes da Técnica

Campo da Técnica

[002] A presente invenção refere-se em geral a reconhecimento de voz automático. Mais especificamente, a invenção refere-se a técnicas para melhorar o reconhecimento automático de voz através do uso de aspectos mais robustos e relevantes do sinal de voz, que incluem informação temporal e padrões derivados de agrupamentos perceptivos e processamento desta informação com o uso de técnicas de aprendizado de máquina.

Descrição da Técnica Relacionada

[003] Informação de percepção de voz é distribuída de maneira não uniforme na frequência, amplitude e tempo. Em todos os aspectos, a voz é altamente variável. A maior parte dos sistemas de reconhecimento de voz automáticos extrai informação em intervalos espaçados uniformemente em uma única escala. Na percepção de voz humana, algumas classes de voz são conhecidas ser distinguidas pela recorrência a características temporais, mas nos sistemas de reconhecimento de voz típicos do estado da técnica os

aspectos temporais da voz não são completamente explorados.

[004] A maior parte dos sistemas de reconhecimento de voz automáticos do estado da técnica inclui um processo que extrai informação do sinal de voz em escalas de tempo uniformes (tipicamente 10 a 15 milissegundos) usando quadros de análise de duração curta uniforme (tipicamente 20 a 30 milissegundos). A classificação da voz baseada em um único vetor de observação de curto prazo não é confiável porque o sinal de voz é altamente dinâmico e constantemente em transição uma vez que vários sons de voz são produzidos. Na verdade, padrões de maior prazo têm que ser empregados para criar sistemas utilizáveis.

[005] Um método conhecido na técnica, que torna padrões de maior prazo disponíveis é reter uma memória de uma quantidade de vetores de observação de curto prazo que em seguida são apresentados simultaneamente para um classificador de voz. Os classificadores usados com esta abordagem são frequentemente redes neurais artificiais ou padrões de correlação. Embora reter uma memória de vetores de observação de curto prazo traga resultados melhorados, existem diversos problemas remanescentes.

[006] Primeiro, a amostragem de escala de tempo uniforme, comum a todos os métodos baseados em quadro, não é síncrona com o sinal de voz. Portanto o relacionamento de eventos da voz e quadros de observação é randômico. Isto resulta em variabilidade aumentada das características extraídas e uma quantização de detalhes temporais.

[007] A seguir, a extração baseada em quadros de análise uniformes não é ótima. A informação usada para percepção humana dos sons da voz ocorre em muitas escalas de tempo diferentes. Por exemplo, o irromper explosivo de um som de "t" falado pode ser tão pequeno quanto uns poucos milissegundos de duração enquanto que uma vogal pode ser sustentada por mais do que um segundo. Uma se-

quência de muitas observações de curto prazo não apresenta a mesma informação que uma informação de longo prazo e vice-versa.

[008] Alguns aspectos da voz são altamente variáveis na dimensão temporal. Por exemplo, a extensão que uma vogal é sustentada depende do orador, da taxa de voz, de se a vogal está em uma sílaba acentuada ou não, de onde na sentença se encontra a palavra que contém a sílaba. Esta variabilidade temporal faz com que a informação de voz se mova para diferentes quadros de observação relativa, aumentando significativamente a variabilidade dos valores extraídos para diferentes exemplos da mesma classe de voz e tornando difícil a detecção de padrões com significado na memória.

[009] Adicionalmente, sistemas baseados em quadro tratam tipicamente todos os quadros com importância igual. Ao contrário, a percepção humana usa as partes do sinal que têm a melhora relação sinal para ruído e que contém as características mais relevantes e confiáveis para fazer as distinções exigidas.

[010] A maior parte dos sistemas de reconhecimento de voz automáticos do estado da técnica incorpora Modelos Markov Ocultos. Os Modelos Markov Ocultos são máquinas de estado estocásticas. Os Modelos Markov Ocultos mapeiam probabilidades estimadas de classe dos vetores de observação em sequências prováveis de produções de classe ocultas (não observadas). Usando os Modelos Markov Ocultos, o problema de variabilidade temporal mencionado acima é tratado permitindo que cada estado de não emissão mude para ele próprio. Usando estados de autotransição a variabilidade temporal é "absorvida". Infelizmente, a menos que a abordagem seja modificada para extrair explicitamente a informação de duração, a abordagem remove informação temporal tanto não desejada como desejada. Os relacionamentos temporais dos eventos de voz trazem informação significativa para percepção dos sons de voz particularmente na discriminação de oclusivas, fri-

cativas e africadas. Além disso, a estimativa robusta de probabilidades de classe requer grandes quantidades de dados de treinamento. Quando as condições de uso diferem das condições de treinamento, as estimativas de probabilidade se tornam muito imprecisas levando a reconhecimento inferior.

[011] As características usadas pela maior parte dos sistemas de reconhecimento de voz automáticos do estado da técnica são primariamente derivadas a partir de perfis espectrais de curto prazo. Esta abordagem é usada frequentemente porque muitos sons de voz têm alguns picos de frequência característicos chamados formantes. Uma abordagem muito diferente empregada por outros sistemas atuais é focar em trajetórias de longo prazo das bandas de frequência. Em um método chamado TRAPs (Padrões Temporais) sons de voz são modelados como as trajetórias médias de longo prazo (~1 s) de exemplos de sons. A classificação é executada baseada na correlação dos pacotes de sinal de voz com cada um dos modelos TRAP. Algumas versões desta abordagem têm resultados reportados comparáveis aos métodos espectrais de curto prazo. Estes resultados mostram que informação útil para identificar os sons de voz está espalhada pelo tempo além dos limites dos segmentos do fonema. Devido à média e janela usadas no método, informação próxima ao centro do TRAP é enfatizada sobre informação mais distante. OsTRAPs capturam tendências brutas mas não capturam detalhes temporais.

[012] Ainda outra abordagem alternativa a extração de característica baseada em quadro é segmentar a voz na localização de certas condições de sinal detectáveis chamadas "eventos". É considerado que cada parte de segmento tem uma única identidade de classe. Usualmente o alinhamento temporal com um modelo é executado pela deflexão dinâmica do tempo, que permite que sejam projetadas as trajetórias das características dentro de uma escala de tempo comum.

Então, na escala de tempo defletida a trajetória da característica é reamostrada e correlacionada com um padrão ou usada como observação para um Modelo Markov Oculto. O processo de deflexão dinâmica do tempo remove muito da variabilidade do tempo dos segmentos de voz. Entretanto, descobrir eventos de segmentação confiáveis apresenta um desafio para métodos baseados em eventos. Inserções e deleções de eventos resultam em desalinhamentos catastróficos.

[013] Claramente existe uma necessidade na área por técnicas melhoradas para aumentar a eficiência e efetividade de reconhecimento de voz automático.

[014] A percepção humana da voz se baseia, em parte significativa, na temporização relativa dos eventos no sinal de voz. Os indicadores para a percepção da voz ocorrem sobre várias escalas de tempo e podem ser deslocadas no tempo a partir da própria percepção. Mudar os relacionamentos temporais de eventos de voz pode mudar a percepção da voz. Isto é demonstrado em B. Repp, e outros, *Perceptual Integration of Acoustic Cues for Stop, Fricative, and Affricative Manner*, Journal of Experimental Psychology: Human Perception and performance 1978, Vol. 4, Num. 4, 621-637, através de experimentos perceptivos onde as durações do silêncio e fricção foram manipuladas. Um destes experimentos introduziu um pequeno intervalo de silêncio entre as palavras "Say" "Shop", que faz com que os ouvintes ouçam "Say Chop". Outro exemplo de como a temporização relativa de eventos influencia a percepção é referenciada como tempo de começo de voz, comumente abreviado VOT. VOT é a extensão de tempo que passa de quando uma parada é liberada até quando a vibração das cordas vocais começa. VOT é um importante indicador na distinção de várias consoantes de parada. A importância da temporização também deriva da variabilidade da duração de fenômenos de voz. Alguns fenômenos de voz perceptíveis são muito breves enquanto que outros são bastante lon-

gos. Por exemplo, o corpus de escritos TIMIT de vozes em Inglês transcritas fonemicamente tem segmentos de interrupção de rajada com durações de menos do que 5 milissegundos, enquanto alguns segmentos de vogais duram mais do que 500 milissegundos.

[015] Embora temporizações relativas de eventos sejam indicadores importantes para a percepção, os métodos mais comuns de extração de características não são sensíveis à temporização de eventos de voz. Quase todas as aplicações de reconhecimento de voz e orador atuais extraem características através da utilização de uma abordagem de segmentação de sinal baseada em quadros de análise de extensão fixa escalonado à frente em tempo por uma dimensão de escala fixa. Devido a estes quadros de análise terem tamanho fixo, os mesmos são quase sempre significativamente menores ou significativamente mais longos do que as extensões dos fenômenos perceptivos que eles tentam capturar.

[016] Embora fácil de implementar, a abordagem comum torna a extração de características sujeita ao relacionamento entre o sinal e o ponto de início do primeiro quadro e ao relacionamento arbitrário entre a dimensão do quadro de análise e a escala de tempo de vários fenômenos de voz. Um sistema de reconhecimento de voz baseado em quadro descrito em S. Basu, e outros, *Time shift invariant speech recognition*, ICSLP98, é baseado em quadros de vinte e cinco milissegundos escalonados por dez milissegundos, deslocamentos no relacionamento inicial do sinal e do primeiro quadro de menos do que dez milissegundos provocaram "modificações significativas das estimativas espectrais e [coeficientes cepstrais de frequência-mel] produzidos pelo adiantamento que por sua vez resulta em variações de até [dez porcento] de taxa de erro de palavras no mesmo banco de dados".

[017] Existem muitas fontes de variação nos sinais de voz: tais como a extensão do trato vocal do orador, acentuação, velocidade da

voz, saúde e estado emocional, bem como ruído de fundo, etc. Entretanto, a variação reportada por Basu e outros é inteiramente devida ao uso de um método de extração de característica no qual a dimensão do quadro e o alinhamento do quadro têm relacionamentos arbitrários com o sinal. A Patente Norte Americana de Número U.S. 5.956.671 (depositada em 4 de junho de 1997) para Ittycheriah e outros, descreveu técnicas voltadas para reduzir variabilidade de características provocadas pelo relacionamento arbitrário entre quadros de análise e sinal de voz. Um aspecto de sua invenção expande a variabilidade do conjunto de treinamento sujeitando múltiplas versões do sinal deslocadas no tempo a processo de análise de quadro fixo como exemplos de treinamento separados. Eles também descrevem uma técnica usada no tempo de reconhecimento onde os valores de características são computados pela ponderação dos resultados de análise de quadro fixo a múltiplas versões atrasadas do sinal.

[018] Estas técnicas não mitigam completamente os problemas provocados pela extração de características usando quadros fixos e escalas de tempo fixas. Além disso, expandir a quantidade de exemplos aumenta o tempo de treinamento e incorpora variabilidade adicional no modelo que não está presente no sinal de voz original. Ponderações deslocadas no tempo aumentam a complexidade e podem "ponderar para fora" algumas características de voz relevantes perceptivamente.

[019] Na Patente Norte Americana de Número U.S. 6.470.311 (depositada em 15 de outubro de 1999), para Moncur, um método de segmentação síncrona de afastamento de voz sonora baseado nos cruzamentos zero positivos da saída de um filtro passa banda com uma frequência central aproximadamente igual ao afastamento endereça parcialmente a sincronização. Voz não sonora é segmentada com o uso do período médio do afastamento computado sobre algum quadro de tempo não especificado. Deve ser observado que condições de baixo

sinal para ruído e sinais com pequenos deslocamentos de sinal DC são conhecidos por causarem problemas para segmentação baseada em cruzamento de zero. Para sinais de voz de alta qualidade, a abordagem de Moncur representa uma melhoria sobre os métodos de análise de quadros fixos comuns durante a voz sonora. Infelizmente para voz não sonora a abordagem reverte para intervalos de tempo e quadros fixos arbitrários. O uso de quadros e intervalos de tempo fixos ainda mantém a localização precisa de eventos tais como fechamento e interrupção de rajadas não resolvidas. Além disso, nenhuma solução é fornecida para voz sussurrada.

[020] Claramente é necessária uma solução que extraia características sincronamente com os eventos do próprio sinal de voz em vez de através de quadros uniformes fixos que têm relacionamentos arbitrários e mutáveis com os fenômenos de voz. A técnica de segmentação deve ser aplicada ao sinal inteiro tanto para voz sonora quanto não sonora. Adicionalmente, a análise de voz deve ser executada sobre escalas de tempo apropriadas para cada um dos tipos de eventos particulares que são detectados.

[021] O mecanismo de reconhecimento de voz automático típico atual espera por um silêncio detectado para analisar e produzir saída porque isto permite uma segmentação natural e, portanto resulta em maior precisão devido ao contexto aumentado. Esperar até o fim de uma voz pode fazer com que a saída seja atrasada em algo entre cinco a vinte e cinco segundos. Quando uma aplicação tem que produzir uma saída em tempo próximo a real, como é exigido por aplicações tais como produção automática de legendas (closed caption) em difusão de televisão, a segmentação menor deve reduzir o contexto disponível para análise, e é esperada e produzida menor precisão. Para estes tipos de aplicações, o que é necessário é alta precisão com baixa latência.

Sumário da Invenção

[022] Algumas modalidades da invenção referem-se à aprendizagem automática de detectores e classificadores para reconhecimento de voz. Mais particularmente, esta invenção é dirigida para a aprendizagem automática de detectores e classificadores que se concentram nos aspectos mais relevantes e robustos do sinal de voz para as tarefas particulares de detecção ou classificação em questão.

[023] Algumas modalidades da invenção envolvem extração de picos ou eventos de sinal de voz que indiquem aspectos notáveis do sinal. Estas modalidades também envolvem capturar os relacionamentos temporais entre os eventos. Nas modalidades preferenciais, um esquema de classificadores ponderados é usado para extrair eventos. Algumas modalidades da invenção envolvem construir o esquema de classificadores ponderados para uso em mecanismo de reconhecimento de voz automático. Algumas modalidades da invenção envolvem detectar sequências de eventos em vez de, ou adicionalmente a, detectar eventos individuais. Em algumas modalidades da invenção, detectores baseados em indicadores alternativos são desenvolvidos.

[024] Em algumas modalidades da invenção, algoritmos de reforço adaptativo são usados para aumentar o desempenho do reconhecimento. Algumas modalidades da invenção incluem um processo para reduzir a complexidade de conjuntos criados por algoritmos de reforço adaptativos.

[025] Em algumas modalidades da invenção, um método para criar automaticamente cascatas de detectores baseados em evento supera os problemas de aprendizado de conjuntos de treinamento altamente desbalanceados ou aprendizagem para detectar objetos raros. A cascata de detectores resultante proporciona detecção eficiente de objetos raros pela eliminação da maior parte de exemplos negativos nos estágios iniciais.

[026] Em algumas modalidades da invenção, um processo de classificar voz em grupos perceptivos é executado. O processo então remove ambiguidades entre percepções alternativas.

[027] Algumas modalidades da invenção envolvem segmentar um sinal de voz em localizações importantes perceptivamente. Isto fornece um meio para não apenas extrair temporizações importantes perceptivamente, mas também sincronizar a análise do sinal com eventos de voz, deste modo evitando todos os problemas de análise de quadro fixo assíncrono. O método primeiro executa uma pré-segmentação com o uso de filtros de baixa complexidade baseados em certos aspectos da percepção humana e nos fenômenos de voz que os mesmos são projetados para detectar. Estes filtros detectam as localizações de padrões perceptíveis que indicam começo, término, rajadas, pulsos glotais, e outros eventos de sinal de voz significativos. Os eventos de pré-segmentação definem intervalos que são usados para sincronizar certas computações de características. Os padrões de características que têm sido extraídos sincronamente são processados adicionalmente para criar características sobre escalas de tempo mais longas e para detectar eventos perceptivos de nível ainda mais alto tais como fronteiras de fonema, núcleos de sílabas, etc.

[028] Preferencialmente, um sistema de reconhecimento de voz de alto nível usa todas estas técnicas. Em algumas modalidades da invenção, é usada uma pluralidade de métodos em um sistema para reconhecimento de voz automático. O sistema recebe uma entrada de voz, aplica um ou mais dos meios de processamento a entrada de voz, decide qual meio de processamento é mais correto, e fornece uma sequência de texto resultante. Nas modalidades atualmente preferenciais da invenção, o sistema de reconhecimento de voz automático é usado em criação de legendas (closed captioning) de televisão em tempo real e ambientes de detecção de palavra. Outras modalidades podem incluir

virtualmente qualquer forma de transcrição de voz, que incluem legendagem ou transcrição de encontros ou conferência telefônica, ditado em tempo real ou conversão oral de mensagens telefônicas para a forma escrita. Algumas modalidades da invenção envolvem processar sinais de voz com o uso de n-séries paralelas de mecanismos de reconhecimento de voz automáticos em modo de rajada sobreposta temporariamente para reduzir a latência. Algumas modalidades da invenção envolvem inserção automática de sinais de pontuação em um texto não pontuado.

Breve Descrição das Figuras

[029] A figura 1 ilustra um exemplo de um fluxo de trabalho para construção de um esquema de classificadores ponderados para uso em um módulo de processamento de um mecanismo de reconhecimento de voz automático de acordo com algumas modalidades da invenção;

a figura 2 ilustra um fluxo de trabalho para identificar automaticamente regiões em uma pluralidade de sinais de voz que contém eventos de acordo com algumas modalidades da invenção;

a figura 3A ilustra os relacionamentos de tempo de eventos de acordo com algumas modalidades da invenção;

a figura 3B ilustra a contagem de eventos que ocorrem dentro das unidades da grade de tempo de acordo com algumas modalidades da invenção;

a figura 3C ilustra a estrutura de um mapa de soma baseado em eventos de acordo com algumas modalidades da invenção;

a figura 4 ilustra um fluxo de trabalho 400 para criar uma cascata de detectores de acordo com algumas modalidades da invenção;

a figura 5 ilustra um exemplo de uma região que contém eventos de todos os exemplos positivos de acordo com algumas modalidades da invenção;

a figura 6A ilustra outro exemplo de uma região no espaço de característica de tempo que contém eventos de todos os exemplos positivos de acordo com algumas modalidades da invenção;

a figura 6B ilustra uma região não alinhada que contém eventos de todos os exemplos positivos de acordo com algumas modalidades da invenção;

a figura 6C ilustra um exemplo de uma região não retangular que contém eventos de todos os exemplos positivos de acordo com algumas modalidades da invenção;

a figura 7 ilustra o relacionamento da fronteira geométrica máxima para as fronteiras mais apertadas e mais folgadas em uma projeção de uma região de acordo com algumas modalidades da invenção;

a figura 8A ilustra uma representação de um sistema automático de voz para texto de acordo com algumas modalidades da invenção;

a figura 8B ilustra uma representação de um sistema automático de voz para texto de acordo com algumas modalidades da invenção;

a figura 8C ilustra uma representação de um sistema para reconhecimento de evento e detecção de palavras de acordo com algumas modalidades da invenção;

a figura 9 ilustra um exemplo de segmentações de um sinal de voz de acordo com algumas modalidades da invenção;

a figura 10 ilustra uma fórmula de contraste perceptivo usada para computar mudança perceptiva de acordo com algumas modalidades da invenção;

a figura 11A ilustra uma memória de fila circular de acordo com algumas modalidades da invenção;

a figura 11B ilustra uma memória de fila circular atualizada de acordo com algumas modalidades da invenção;

a figura 11C ilustra uma memória de fila circular atualizada de acordo com algumas modalidades da invenção;

a figura 12 ilustra uma fila circular segmentada para manter somas correntes de acordo com algumas modalidades da invenção;

a figura 13 ilustra uma fila circular segmentada de acordo com algumas modalidades da invenção;

a figura 14 ilustra uma representação de uma saída do detector de pulso glotal em um pequeno segmento de voz sonora de acordo com algumas modalidades da invenção;

a figura 15 ilustra uma representação de um detector de núcleos silábicos de acordo com algumas modalidades da invenção;

a figura 16 ilustra um fluxo de trabalho para executar extração de formante de acordo com algumas modalidades da invenção;

a figura 17 ilustra um fluxo de trabalho para executar extração harmônica de acordo com algumas modalidades da invenção;

a figura 18 ilustra uma representação de dois mecanismos de processamento consecutivos, se sobrepondo no tempo, operando em uma sequência de dicções de acordo com algumas modalidades da invenção; e

a figura 19 ilustra um sistema de voz para texto que inclui um pontuador automático para algumas modalidades da invenção.

Descrição Detalhada da Invenção

[030] A invenção refere-se à aprendizagem automática de detectores e classificadores para reconhecimento de voz. Mais particularmente, esta invenção é direcionada para a aprendizagem automática de detectores e classificadores que se concentram nos aspectos mais relevantes e robustos do sinal de voz, incluindo informação temporal para as tarefas particulares de detecção ou classificação em questão.

[031] Nas modalidades atualmente preferenciais da invenção, o sistema de reconhecimento de voz automático é usado em criação de

legendas (closed captioning) de televisão em tempo real e ambientes de detecção de palavra.

[032] Embora o reconhecimento de voz automático tenha melhorado ao longo dos anos o mesmo ainda não se aproxima do desempenho humano. Níveis de ruído que não provocam nenhuma dificuldade para ouvintes humanos podem frequentemente tornar inutilizáveis sistemas de reconhecimento de voz automáticos no estado da técnica. Melhorias na precisão são provenientes, acima de tudo, ao custo de acréscimo de tempo de processamento e complexidade computacional. Em parte significativa estas dificuldades provém do fato de que a informação usada pelos humanos para percepção de voz é distribuída de maneira não uniforme na frequência, amplitude e tempo. A maior parte dos sistemas de reconhecimento de voz trata todos os pontos no tempo como igualmente relevantes para a percepção da voz e faz todas as classes de determinações baseadas no mesmo conjunto de características. Os seres humanos, por outro lado, parecem ser capazes de selecionar aqueles aspectos do sinal de voz que são mais relevantes e robustos para fazer as distinções necessárias para percepção.

[033] Os receptores neurais no ouvido convertem o sinal acústico em padrões temporais de picos relacionados às suas características de amplitude dinâmica e distribuição de frequência. Os padrões de picos temporais codificam a informação e a comunicam para os neurônios do cérebro para processamento adicional. Os neurônios e sinapses que formam as unidades computacionais do cérebro usam padrões de picos para codificar e comunicar informação um para o outro. A eficiência e efetividade do reconhecimento de padrão do maquinário neural humano são excepcionais. A codificação de picos cria uma representação muito esparsa do sinal. Inspirada por certos aspectos da percepção humana, a presente invenção codifica informação extraída do sinal de voz como picos, referenciados neste documento como "eventos".

[034] Nas modalidades atualmente preferenciais da invenção, a extração baseada em eventos se concentra nos aspectos notáveis do sinal e captura os relacionamentos temporais destes aspectos. Um exemplo de um tipo de evento seriam picos em pacotes de energia de bandas de frequência passantes. Os picos são as localizações no sinal de voz onde a energia da voz em cada banda é mais forte contra o ruído de fundo. A distância temporal entre picos e a sequência temporal de eventos são fortemente relacionadas ao que está sendo falado. A extração de evento não é limitada a achar os picos de pacotes de filtros passa banda. Outros eventos incluem eventos de começo e deslocamento gerados através de análise mais complexa do sinal que inclui a saída de detectores de subpadrão. Classificadores e detectores baseados em qualquer método conhecido podem ser incorporados dentro de padrões de evento fazendo com que os mesmos disparem quando as condições para as quais os mesmos foram designados são detectadas.

Construindo Detectores e Classificadores Automáticos Relevantes

[035] Como usado aqui, o termo "classificadores" refere-se a um método e aparelho que designa rótulos de classe a vetores de características, eventos, e/ou sequências de eventos. Detectores são classificadores, os quais designam rótulos de classe de "presente" ou "ausente" a cada vetor de característica, evento, e/ou sequência de eventos.

[036] Classificadores de átona são funções de decisão que executam melhor do que o acaso. Conjuntos classificadores são formados através da combinação de resultados de múltiplos classificadores de átona. Reforço é um método conhecido na técnica para construir automaticamente conjuntos qualificadores através da seleção e ponderação de classificadores de átona de modo que a decisão do conjunto seja melhor do que as decisões de qualquer um dos classificadores de átona. A seleção é feita avaliando iterativamente cada classificador de

átona de um conjunto relativamente grande de classificadores de átona e escolhendo aquele que tem o melhor desempenho em uma distribuição ponderada dos exemplos de treinamento rotulados. O classificador de átona selecionado é adicionado ao conjunto e é designado um peso à sua decisão baseado em sua taxa de erro. Os pesos de distribuição são então ajustados para enfatizar os erros feitos pelo conjunto e a próxima iteração é iniciada. Devido aos exemplos que não foram classificados corretamente serem enfatizados na distribuição, os classificadores de átona que tendem a corrigir os erros do conjunto são adicionados em etapas subsequentes e as decisões globais do conjunto são melhoradas.

[037] Reforço tem mostrado que gera classificadores com boas características de generalização. Os classificadores de átona podem tomar qualquer forma desde que seu desempenho seja melhor do que ao acaso.

[038] Um método para executar classificação de padrão temporal é amostrar as trajetórias da característica em múltiplos intervalos fixos e apresentar todos os pontos de característica-tempo como características individuais. Tipicamente, uma quantidade fixa de pontos característica-tempo é usada para classificação. Com uma quantidade fixa de pontos característica-tempo, a correspondência entre informação em um exemplo e aquela de outro exemplo é estabelecida pela definição do vetor de característica.

[039] De acordo com as modalidades preferenciais da invenção atualmente, é usada uma abordagem diferente. Devido à amostragem uniforme de trajetórias de característica poderem perder detalhes que ocorrem entre amostras e a amostragem uniforme criar muitas amostras que contêm pouca informação discriminada, a invenção em vez disso amostra trajetórias de característica relativa a eventos. Eventos são os pontos nas trajetórias onde é localizada informação significati-

va. Extração baseada em eventos cria uma representação esparsa do sinal. Esta abordagem requer modificação do método para definir classificadores de átona usados tipicamente em outros contextos, tais como processamento de imagem, porque exemplos de uma dada classe podem ter zero, um, ou mais do que um evento de um dado tipo, portanto é necessário um método para estabelecer correspondência entre informação em um exemplo e informação em outro exemplo.

[040] Valores de características, eventos e padrões de eventos podem fornecer evidência que seja consistente com a classe-alvo do detector ou podem fornecer evidência contrária. Os tipos de eventos, e os relacionamentos temporais entre eventos, representam uma parte significativa da evidência a favor ou contra uma detecção de classe-alvo. Infelizmente, a correspondência exata entre padrões de eventos em exemplos diferentes da mesma voz não ocorrem. Além disso, o ruído pode provocar eventos espúrios ou faltantes, e a velocidade da voz pode provocar variação temporal nas sequências de eventos. Usualmente as técnicas de aprendizagem de máquina são projetadas para utilizar vetores de característica de extensão fixa. Com vetores de característica de extensão fixa, cada exemplo de treinamento positivo e negativo sempre tem um valor para toda característica e a correspondência entre valores de característica para cada exemplo é achada na mesma localização indexada no vetor de característica. Diferente dos valores em vetores de característica de extensão fixa, eventos e padrões de eventos podem existir ou não e podem ter relacionamentos temporais de alguma forma diferentes um com o outro fazendo com que seja difícil determinar quais eventos de um exemplo correspondem, a um evento em outro exemplo.

[041] A invenção define métodos através dos quais a correspondência de eventos e padrões de eventos entre exemplos pode ser determinada, de modo que informação temporal possa ser explorada pa-

ra criar detectores de áfona para conjunto de aprendizes reforçados.

[042] Nas modalidades da invenção preferenciais atualmente uma origem temporal é associada com um evento de um certo tipo, e as origens temporais de todos os exemplos são alinhadas. As variações temporais de eventos que representam um certo aspecto de voz são limitadas por um intervalo definido relativo à origem temporal. Para um dado intervalo, se existe uma diferença na consistência com a qual os eventos (de certo tipo) caem dentro do intervalo para a classe positiva e a classe negativa, a diferença pode ser explorada para criar um detector de átona. Em algumas modalidades desta invenção, exemplos são alinhados baseados na localização de seus eventos de núcleo silábico. Em algumas modalidades desta invenção, conjuntos de dois ou mais eventos são alinhados com respeito a um dos eventos dentro de cada conjunto.

[043] Para fazer um detector de átona utilizável baseado em informação afirmativa associada com eventos, os intervalos que definem o detector de átona têm que conter eventos em sua maioria exemplos positivos e têm que não conter eventos em uma maioria de exemplos negativos. Estes intervalos podem ser sistematicamente determinados através da avaliação de todos os intervalos que contém eventos de uma maioria de exemplos positivos. Primeiro, os exemplos são trazidos para correspondência temporal geral através de alinhamento baseado em um evento comum particular. Opcionalmente, exemplos de diferentes durações gerais podem ser colocados em escala para ter uma extensão comum. Os intervalos consistentes podem ser descobertos eficientemente através de primeiro, para todos os exemplos, arrumar os eventos de sensores diferentes (por exemplo, sensores de banda de frequência) em dois espaços dimensionais e gravar a soma acumulada da quantidade ponderada de eventos acima e a esquerda de cada evento. Em seguida a quantidade de eventos dentro de qual-

quer intervalo retangular pode ser determinada através de diferenças simples nas contas ponderadas acumuladas. Detectores de átona baseados em cada intervalo que contém eventos para a maioria dos exemplos são avaliados e o melhor detector para a distribuição ponderada corrente é retido. O detector composto é avaliado no conjunto de treinamento inteiro e os pesos de distribuição são ajustados para os erros feitos.

[044] Classificadores de átona são adicionados de acordo com o processo acima até que o detector de desempenho esteja perfeito nas amostras de treinamento ou que seja alcançada a quantidade máxima de iterações.

[045] A figura 1 ilustra um exemplo de um fluxo de trabalho 100 para construir um esquema de classificadores ponderados para uso em um módulo de processamento de um mecanismo de reconhecimento de voz automático. Nas modalidades da invenção preferenciais atualmente, o esquema de classificação ponderada é usado no módulo de classificação de um mecanismo de reconhecimento de voz automático, como explicado abaixo em conexão com a figura 9. O fluxo de trabalho 100 da figura 1 começa armazenando uma pluralidade de sinais de voz como um conjunto de treinamento 101 e em seguida extrai padrões de eventos do conjunto de treinamento 102, em que os ditos padrões de evento compreendem aspectos característicos dos sinais de voz. A seguir, uma amostra de sinais de voz com padrões de evento correspondentes é acessada 103 e alinhada com base na localização temporal de onde o evento ocorreu dentro do sinal de voz 104. Cada sinal é em seguida colocado em escala opcionalmente para uma duração temporal comum 105.

[046] Uma vez que os sinais extraídos sejam colocados em escala para uma duração comum com localizações de evento comuns, uma pluralidade de detectores de átona é aplicada aos sinais e a efetivida-

de de cada classificador de átona é testada em sua habilidade de detectar os eventos 106. Baseado na efetividade medida, os classificadores de átona são ponderados, com aqueles que executam bem recebendo um coeficiente alto e aqueles que executaram de forma deficiente recebendo um coeficiente baixo 107.

[047] A seguir a efetividade do esquema de ponderação é testada para determinar se a ponderação reconhece adequadamente eventos no conjunto de treinamento baseado em um limite de efetividade predeterminado 108. O fluxo de trabalho faz uma consulta se a ponderação reconhece adequadamente os eventos 109. Se o esquema de ponderação executa adequadamente, o fluxo de trabalho 100 armazena o esquema de ponderação e termina 110. Por outro lado, se o esquema de ponderação não executa adequadamente, são adicionados classificadores de átona ao grupo de classificadores de átona aplicado previamente 111, e o fluxo de trabalho reitera até que o nível limite de efetividade seja alcançado.

[048] Os padrões de evento de diferentes exemplos de uma dada voz têm alguma similaridade, entretanto, não ocorre correspondência exata de eventos entre quaisquer dois exemplos de voz. Se é dado um tempo de referência comum a eventos de exemplos diferentes, tal como sendo feitos relativos aos centros de sílabas, os eventos correspondentes de diferentes exemplos de uma dada voz ocorrerão dentro de uma região no plano tempo-sensor. Voz é altamente variável e a informação mais útil para percepção é distribuída não uniformemente em frequência, amplitude, tempo e escala de tempo. Portanto, especificar regiões no plano tempo-sensor que contêm eventos que contribuem com certa informação perceptiva não pode ser feito efetivamente usando uma única escala ou forma constante. Entretanto, avaliar completamente todas as possíveis posições, formas e escalas de regiões que possam conter coleções ou eventos correspondentes relevantes

pode ser computacionalmente inviável. Portanto, é definido um processo que identifica automaticamente regiões de eventos correspondentes úteis para percepção de voz.

[049] Primeiros eventos de uma pluralidade de exemplos de treinamento positivos são feitos relativos a uma referência de tempo comum, tal como centros de sílabas e os eventos são projetados no plano tempo-trajetória. Opcionalmente, antes da projeção os padrões podem ser colocados em escala de modo que sua duração seja igual a 1. Regiões no plano tempo-trajetória que contém eventos de uma maioria de exemplos positivos são retidas como grupos potenciais de eventos correspondentes. Uma lista destas regiões é formada e usada para todas as etapas subsequentes de criação de detectores de átona.

[050] A figura 2 ilustra um exemplo de um fluxo de trabalho 200 para identificar regiões automaticamente em uma pluralidade de sinais de voz que contêm padrões de evento de acordo com algumas modalidades da invenção. O fluxo de trabalho 200 começa alinhando um grupo de sinais de voz de um conjunto de treinamento de sinais de voz relativos a um eixo de tempo comum 201. A seguir, o fluxo de trabalho 200 opcionalmente coloca em escala a duração de cada sinal de voz individual no grupo para uma unidade de duração de tempo comum 202 e que projeta centros de sílaba dos sinais de voz e os centros de eventos dos sinais de voz no eixo do tempo comum 203. Finalmente, as regiões do eixo do tempo que têm uma concentração alta de centros de silaba e centros de evento são identificadas como regiões que contém padrões de evento 204.

[051] Adicionalmente às técnicas reveladas para identificar regiões que têm uma alta concentração de eventos, a invenção também envolve diversas técnicas que são empregadas para rejeitar regiões que são pouco prováveis de resultar em detectores de átonas robustos que incluem, mas não são limitadas a mapeamento de integração de

evento, aplicação de restrições de densidade de exemplo, rejeição de regiões redundantes e combinações dos mesmos.

Mapeamento de Integração de Eventos

[052] Em algumas modalidades da invenção, um processo de mapeamento de integração de eventos é empregado para rejeitar regiões que não são prováveis de resultar em detectores de átona úteis.

[053] Uma técnica conhecida no campo de processamento de imagem que permite a computação rápida da soma de valores de pixel sobre regiões retangulares é modificada para permitir rejeição rápida de regiões inviáveis com base em contadores de eventos na região. Na técnica de processamento de imagem original a primeira etapa é computar um "mapa de soma" no qual cada célula do mapa corresponde à soma dos valores de pixel na região retangular definida pelo canto naquela célula e o canto diagonalmente oposto na origem. Após este mapa de soma ter sido computado a soma dos pixels de qualquer sub-região retangular da imagem pode ser determinada com duas operações de subtração e uma de adição. A técnica de "mapa de soma" é adaptada para a eliminação rápida de regiões que não podem conter evidência de mais do que uma quantidade especificada de eventos em cada célula da grade de uma grade sobreposta no plano tempo-trajetória. Quando um mapa de soma de contas de eventos de célula de grade é computado então pode ser feita uma determinação da quantidade de eventos em qualquer região retangular usando apenas duas operações de subtração e uma de adição. Conhecer a quantidade de eventos na região não é equivalente a conhecer a quantidade de exemplos na região, mas isto estabelece o limite superior. Portanto qualquer região que não tenha um contador de eventos maior ou igual à quantidade requerida de exemplos possivelmente não pode conter a quantidade requerida de exemplos.

[054] As figuras 3A a 3C ilustram a estrutura de um mapa de soma

baseado em eventos de acordo com algumas modalidades da invenção. Na figura 3A é representado um padrão de eventos no plano tempo-trajetória. Na figura 3B são determinados os contadores de eventos que ocorrem dentro dos limites de uma grade sobreposta. Na figura 3C é conhecido um mapa de soma onde cada célula contém a soma dos contadores na região retangular que tem a origem como um canto e a célula como o canto diagonalmente oposto. Para determinar a quantidade de eventos nas quatro células centrais da figura 3C, a partir do valor na célula superior direita da região em questão, neste caso "7", o valor da região não incluída à esquerda é subtraído, neste caso "3", como na região não incluída abaixo, neste caso "4", e a região subtraída acima na interseção das duas regiões subtraídas é adicionada de volta, neste caso "2". Isto resulta na quantidade de eventos na região, neste caso "2" (7-3-4+2 = 2). O custo computacional de determinar os contadores de evento de uma região de qualquer dimensão ou forma é o mesmo.

Restrição de Densidade de Eventos

[055] Em algumas outras modalidades da invenção, a aplicação de restrições de densidade de eventos é empregada para rejeitar regiões que não são prováveis de resultar em detectores de átona úteis. Por exemplo, restrições de densidade mínima podem ser aplicadas opcionalmente para rejeitar regiões com uma densidade de eventos abaixo de uma quantidade especificada.

Rejeição de Região Redundante

[056] Em algumas modalidades da invenção, regiões redundantes que são improváveis de resultar em detectores de átona úteis são rejeitadas. Regiões que contém outras regiões, mas não adicionam eventos positivos adicionais além daqueles incluídos dentro da região contida não são adicionadas a lista de regiões.

[057] Com referência novamente a figura 2, uma vez que as regiões são identificadas, as mesmas formam restrições que são usadas

para gerar detectores de átona. Os detectores de átona podem consistir em um simples teste para determinar se um dado exemplo tem quaisquer eventos dentro da região ou não, ou podem ser estendidos para incluir restrições adicionais baseadas na extensão de valores de característica dos exemplos positivos que tem eventos dentro da região.

Reconhecimento de Voz Baseado em Sequência de Eventos

[058] As sequências de eventos, em geral, são discriminadores mais poderosos no reconhecimento automático de voz do que eventos individuais dos quais eles são compostos. Algumas modalidades da Invenção envolvem detectar sequências de eventos em vez de, ou adicionalmente a, detectar eventos individuais.

[059] Em algumas modalidades da invenção, uma sequência de eventos é localizada como um ponto no hiperespaço pelo uso dos intervalos (possivelmente colocados em escala) em espaço temporal-sensor como coordenadas. Para entender o conceito, considera-se a sequência de três eventos produzidos por um único sensor, no qual o segundo evento segue o primeiro por duas unidades de tempo e o terceiro segue o segundo por quatro unidades de tempo. A sequência de tempo destes três eventos com respeito a cada um dos outros é representada pelas coordenadas (2, 4). A similaridade das sequências temporais pode ser julgada através da computação de uma função de distância entre os pontos projetados. Por exemplo, a distância Euclideana pode ser usada para este propósito. Para acessar quais sequências podem aparecer consistentemente (ou não) nos exemplos, as sequências de eventos de um exemplo positivo são projetadas como já descritas para formar um conjunto de pontos-padrão que representam as sequências possíveis que podem ser associadas com os exemplos positivos. Um ponto-padrão é definido baseado nas coordenadas de cada um dos pontos do primeiro exemplo e cada contador associado de ponto-padrão é ajustado para 1. As sequências de eventos do res-

tante de eventos positivos são projetadas dento dos pontos de hiperespaço com o uso de seus intervalos como coordenadas de maneira semelhante ao primeiro exemplo. Conforme cada ponto da sequência é gerado o mesmo é associado com o ponto-padrão mais próximo. O ponto de sequência é adicionado a uma lista associada com aquele ponto-padrão e o contador de pontos-padrão é incrementado de 1. As coordenadas de ponto-padrão são então ajustadas para se tornarem os valores médios das coordenadas de seus pontos de exemplo associados. Após todos os exemplos terem sido processados, os pontos-padrão com contadores altos representam sequências de eventos que são altamente associadas com a classe. As coordenadas dos pontos-padrão representam os centros relativos das regiões com respeito ao primeiro evento na sequência. O tamanho e formas das regiões podem ser determinados pela variação das sequências de exemplos associados. Em algumas modalidades da invenção, pode ser desejável mesclar sequências similares. Candidatos para o consolidador são determinados facilmente por sua distância no hiperespaço projetado.

[060] Em algumas modalidades da invenção o processo descobre combinações de regiões que detectam sequências de eventos que frequentemente ocorrem juntamente com a classe-alvo. A utilidade destes detectores de átona depende da co-ocorrência ser menos frequente quando a classe-alvo não está presente.

[061] O processo descrito neste documento envolve um processo para descobrir sequências de eventos que fornecem evidência afirmativa da classe positiva. Evidência em contrário também tem valor. Para descobrir evidência contrária, o processo descrito acima é repetido, mas desta vez com os exemplos negativos. Detectores inibidores de átona são formados baseados nas sequências que recorrem nos exemplos negativos com alguma frequência, mas nunca ou raramente ocorrem nos exemplos positivos.

[062] Em algumas modalidades da invenção, conjuntos de detectores de átona podem ser formados através do uso de um algoritmo de reforço adaptativo para manusear conjuntos de treinamento não balanceados ou para resultar em detectores de menor complexidade.

Melhoria de Desempenho Através de Simplificação de Conjuntos Reforçados

[063] Em algumas modalidades da invenção, algoritmos de reforço adaptativos são usados para aumentar o desempenho do reconhecimento. Algoritmos de reforço adaptativos envolvem um processo reiterativo de chamar sequencialmente classificadores de átona, testar estes classificadores e ajustar os coeficientes de ponderação adequadamente. Algoritmos de reforço adaptativos criam conjuntos pela adição de um detector de átona por iteração sem inspecionar a frente e sem correção de pesos anteriores. Como resultado, o conjunto final pode ser mais complexo do que o necessário.

[064] Algumas modalidades da invenção incluem um processo para reduzir a complexidade dos conjuntos criados por algoritmos de reforço adaptativos. De acordo com estas modalidades após o detector alcançar perfeição no conjunto de treinamento ou alcançar uma quantidade máxima de iterações, então um processo de simplificação é executado. O desempenho do detector composto é comparado iterativamente com versões dele próprio em que cada um tem uma diferença de seus detectores de átona removidos. Se remover qualquer um dos detectores de átona melhora a taxa de erro, a remoção que tem o maior ganho é executada, caso contrário, se remover qualquer um dos detectores de átona não provoca aumento na taxa de erro, um destes detectores é removido. O processo continua até que não sejam mais removidos detectores de átona.

[065] Em outras modalidades da invenção, é usado um algoritmo de reforço de programação linear que atualiza todos os pesos do con-

junto conforme são adicionados novos detectores para a construção de conjuntos.

Detecção de Indicador Alternativo

[066] A percepção humana de voz pode se basear em indicadores alternativos quando alguns aspectos do sinal de voz são corrompidos. Igualmente, indicadores alternativos podem ser achados em uma amostra de voz e detectados em um sistema de reconhecimento de voz automático.

[067] Em algumas modalidades da invenção, detectores baseados em indicadores alternativos são desenvolvidos seguindo as etapas mencionadas acima para criar um conjunto detector e então repetindo o processo para fazer os detectores subsequentes com a restrição de que detectores de átona usados pelos detectores criados previamente não podem ser usados para construir os detectores subsequentes. Isto maximizará a independência dos detectores. Detectores de múltiplos indicadores alternativos podem então ser combinados como um conjunto para fazer um detector que seja tolerante a tal variação.

Conversão Automática de Conjuntos Para Detectores em Cascata

[068] A decisão global do conjunto é a soma ponderada dos detectores individuais. Na forma-padrão do conjunto, todos os classificadores de átona têm que ser avaliados para fazer uma determinação de voz. Em algumas modalidades da invenção o conjunto de detectores é convertido em um detector em cascata que reduz a quantidade de detectores de átona que tem que ser avaliados na média. Ordenar os detectores de átona do mais forte para o mais fraco e analisar o relacionamento entre as somas em cada estágio e o resultado final, podem estabelecer limites "precoces" que convertem o conjunto para uma cascata de detectores.

[069] O sincronismo relativo de vários eventos contém informação importante para percepção de voz. Este tipo de informação pode

ser explorado pelo exame de padrões persistentes de eventos correspondentes a partir de múltiplos eventos de uma dada palavra, sílaba, fonema, etc. Esta análise é desafiadora devido à variabilidade em todos os aspectos da voz e pelo fato de que ocorrem diferentes indicadores perceptivos sobre escalas de tempo diferentes.

[070] Entretanto, como explicado neste documento, a maior parte das técnicas de classificação de aprendizagem de máquina é designada para aprender decisões baseadas em vetores de dimensão fixa de informação homogênea. Com extração baseada em evento, os eventos ocorrem ou não de acordo com condições de sinal. Isto significa que um dado exemplo pode ter mais ou menos eventos do que outro exemplo da mesma sílaba, palavra, fonema, etc. A fim de treinar detectores de treinamento com eficácia usando extração baseada em evento, é necessário descobrir quais eventos de um exemplo de uma sílaba, palavra, fonema, etc., correspondem ao mesmo suporte perceptivo em outros exemplos. Posteriormente neste documento, são descritos métodos que localizam automaticamente os limites destes eventos correspondentes.

Métodos e Técnicas Para Usar Exemplos de Treinamento Automaticamente para Descobrir Suporte Relevante e Informação Contrária e Determinar Pesos Para Fazer uma Decisão de Detecção

Cascatas Baseadas em Eventos para Conjuntos de Treinamento Altamente Desbalanceados

[071] Em algumas modalidades da invenção, um método para criar automaticamente cascatas de detectores baseadas em eventos supera os problemas de aprendizagem a partir de conjuntos de treinamento altamente desbalanceados ou de aprendizagem para detectar objetos raros. As cascatas de detectores resultantes fornecem detecção eficiente de objetos raros por eliminação da maioria dos exemplos negativos nos estágios iniciais.

[072] Em algumas modalidades da invenção, criar cascatas de detectores baseadas em eventos envolve criar detectores para palavras específicas que ocorrem raramente. Detectar palavras raras é usado simplesmente para ilustrar a invenção e outras aplicações de detecção ficarão prontamente aparentes para os indivíduos com conhecimentos comuns na técnica tendo o benefício desta descrição. Por exemplo, algumas outras técnicas incluem detecção de classes de voz de subpalavras, tais como sílabas específicas, fonemas específicos, classes amplas de sílabas e classes amplas de fonética. Adicionalmente, a invenção pode ser aplicada a muitas aplicações que não são relacionadas a reconhecimento de voz tais como monitoração de processo industrial, detecção de falha em sistema automotivo e monitoração de equipamento médico.

[073] Conjuntos de treinamento altamente desbalanceados com poucos exemplos positivos e muitos exemplos negativos não são bem manipulados através de técnicas de aprendizagem de máquina com tentativa de minimizar a quantidade de erros. Quando exemplos positivos ocorrem raramente, por exemplo, com uma taxa de 1 em 100.000.000 então um detector que falha em detectar esta ocorrência deve ter uma taxa de erro muito baixa (taxa de erro = 0,00000001). Entretanto, mesmo embora o mesmo tenha uma taxa de erro baixa devido à nunca fazer uma detecção falsa, ele é essencialmente inútil.

[074] Objetos que são membros de uma classe compartilham características cujos valores ficam dentro de certas amplitudes. Deste modo, objetos com características cujos valores fiquem fora destas amplitudes podem ser totalmente rejeitados por não pertencerem a classe. Entretanto, objetos que têm característica cujos valores não estão completamente dentro da amplitude podem ter algumas características cujos valores ficam dentro da amplitude associada com a classe. Portanto, pode ser possível excluir associação a classe para um

objeto se o mesmo tem um único valor de característica fora da amplitude. Consequentemente, em algumas modalidades da invenção, confirmar associação a classe geralmente exige que todos os valores de característica relevantes estejam dentro de amplitudes consistentes com a classe.

[075] Aplicada ao reconhecimento de voz, extração de característica baseada em eventos cria uma representação esparsa que preserva a informação mais relevante para reconhecimento de classes de voz incluindo informação temporal. Um exemplo de um tipo de evento que pode ser extraído é a ocorrência de um pico no pacote de uma certa trajetória de característica. Um pacote de trajetória de característica pode, por exemplo, ser computado na saída produzida quando o sinal de voz é passado através de certo filtro passa banda. Quando muitas destas trajetórias são computadas, os eventos são distribuídos no espaço tempo-trajetória. Toda evidência útil para identidade de classe de palavra é associada com eventos no espaço tempo-trajetória. Quando tempos de eventos são tornados relativos a uma referência de tempo comum, tal como os centros de sílaba, e os eventos de múltiplos exemplos da mesma classe são plotados no espaço tempo-trajetória, são formadas regiões que contêm agrupamentos de eventos relacionados.

[076] As localizações, forma e escala destas regiões que contêm agrupamentos são específicas à classe. Algumas destas regiões ficarão tão fortemente associadas com a classe que todos os eventos positivos da classe terão eventos que ficam dentro da região. Como explicado acima, um objeto que não tenha um evento dentro desta região pode ser rejeitado como sendo membro de outra classe. Vários valores de característica podem ser associados com cada evento. As amplitudes de valores para cada uma das características associadas com eventos de exemplos de classe positiva dentro de uma região formam

intervalos em dimensões adicionais do espaço. Um objeto tem que ter um evento com valores associados dentro da amplitude de toda dimensão de característica relevante para ser aceito como um membro da classe. As características que distinguem um objeto que não é da classe de todos os objetos da classe podem ser diferentes das características que distinguem outro objeto que não é da classe de todos os objetos da classe.

[077] De acordo com algumas modalidades da invenção, estas considerações relevantes podem ser descobertas automaticamente para criar um detector. A figura 4 ilustra um fluxo de trabalho 400 para criar uma cascata de detectores de acordo com algumas modalidades da invenção.

[078] O fluxo de trabalho 400 começa pela inicialização da cascata de detectores para conter zero estágios de detector 401. A seguir, todas as regiões no espaço tempo-trajetória que contém eventos de todos os exemplos de treinamento positivos são identificadas e a quantidade de exemplos negativos que tem eventos dentro de cada região identificada é computada 402.

[079] Em seguida, para cada região que contém eventos de todos os exemplos de treinamento positivo, a definição de regiões pode opcionalmente ser expandida para incluir dimensões de característica adicionais 403. Os limites da região para quaisquer dimensões adicionais são selecionados de modo que os mesmos incluam a amplitude total de valores dos exemplos positivos. Em seguida, exemplos negativos que não incluem valores de característica dentro de todas as amplitudes, assim estabelecidas, são rejeitados e o contador de exemplos negativos incluídos na região é reduzido de acordo 404. As dimensões adicionais, se houver alguma, são escolhidas para minimizar o contador de exemplos negativos incluídos dada a quantidade de dimensões. Isto significa que as dimensões de característica usadas para diferen-

tes regiões são aquelas que discriminam melhor e podem variar de região para região.

[080] Em seguida a região na lista que contém eventos da menor quantidade de exemplos de treinamento negativos é selecionada como um estágio de cascata de detectores 405. Em algumas modalidades da invenção, uma quantidade máxima de estágios de detectores é predeterminada. Adicionalmente, exemplos negativos sem eventos na região selecionada são eliminados de consideração adicional 406.

[081] Em seguida, o fluxo de trabalho faz pesquisa sobre quantos exemplos negativos restam 407. Se não existem mais exemplos negativos restantes, foi criada uma cascata de detectores que executa perfeitamente nos exemplos de treinamento, o fluxo de trabalho 400 fornece o detector 408 e para.

[082] Se não existem menos exemplos negativos do que na iteração anterior não pode ser feita melhoria adicional. Neste caso, o fluxo de trabalho 400 remove o estágio recém-adicionado, transcreve o detector imperfeito 409, e para.

[083] Ao contrário, se existem menos exemplos negativos do que na iteração anterior, então o fluxo de trabalho pesquisa se a quantidade máxima de estágios de detector foi adicionada 410. Se a quantidade máxima de estágios de detector foi adicionada à cascata, então o fluxo de trabalho 400 fornece um detector imperfeito 411 e para.

[084] Se existem exemplos negativos remanescentes e se a quantidade máxima de estágios de detector não foi alcançada, então o fluxo de trabalho 400 reitera e continua a construir a cascata de detectores através da inclusão de estágios adicionais voltando à etapa 402.

[085] Após as cascatas de detectores serem criadas as mesmas são usadas de acordo com o método a seguir. Primeiro, são detectados eventos e dada uma referência comum como feito durante o processo de treinamento. Em seguida, começando do primeiro estágio da

cascata, os eventos na lista são avaliados para determinar se qualquer um está dentro da região. Se quaisquer eventos são considerados dentro da região, os eventos na lista são avaliados por estágios subsequentes desde que pelo menos um evento seja considerado dentro da região considerada por aquele estágio.

[086] Em seguida, se o objeto tem eventos dentro das regiões de todos os estágios da cascata, o objeto é detectado como um membro da classe. Finalmente, se o objeto não tem eventos em qualquer um dos estágios, o mesmo é rejeitado como um membro da classe por aquele estágio e nenhum processamento adicional é executado.

[087] Nestes exemplos, regiões (hiper-) retangulares alinhadas ao eixo foram utilizadas. Em algumas outras modalidades da invenção, outras configurações de limites são usadas, tais como (hiper-) esferas, ou (hiper-) elipses ou misturas das formas de limites em diferentes regiões ou em diferentes dimensões. Além disso, podem ser usadas regiões (hiper-) retangulares que não são alinhadas ao eixo. Esta observação se aplica a todas as referências a detectores de átona.

[088] As figuras 5 a 6C ilustram vários exemplos de projeções de eventos de exemplo de treinamento em um plano de valor de tempo-característica de acordo com algumas modalidades da invenção. A figura 5 ilustra um exemplo de uma região que contém eventos de todos os exemplos positivos. A figura 6A ilustra outro exemplo de uma região que contém eventos de todos os exemplos positivos. A figura 6B ilustra uma região não alinhada que contém eventos de todos os exemplos positivos. A figura 6C ilustra um exemplo de uma região não retangular que contém eventos de todos os exemplos positivos.

Melhorando a Generalização Através da Maximização da Margem Geométrica

[089] O método usado para identificar regiões no plano tempo-trajetória resulta em limites que são ajustados em volta dos eventos de

exemplo de treinamento positivos contidos na região. Quando usado como um detector estes limites ajustados devem rejeitar casos em que os valores que são apenas ligeiramente diferentes dos eventos de exemplo de treinamento nos limites externos da região. Se os limites são expandidos tanto quanto possível sem abranger eventos de exemplos negativos adicionais, o detector será capaz de detectar casos com valores similares, mas além da amplitude de valores de qualquer dos exemplos de treinamento positivos na região. Entretanto, estes limites folgados ao máximo podem provocar detecções falsas de casos com valores que são apenas ligeiramente diferentes dos valores de eventos de exemplo negativo próximos aos limites.

[090] A generalização pode ser melhorada ajustando cada um dos limites na região para maximizar a margem geométrica entre eventos de exemplo positivo detectado e eventos de exemplo negativo rejeitados. Os limites de margem geométrica máximos ficam no meio do caminho entre os limites mínimos ajustados e os limites máximos folgados. Maximizar as margens geométricas proporciona a melhor oportunidade para generalização para casos ocultos em exemplos de treinamento. A figura 7 ilustra o relacionamento do limite geométrico máximo para os limites maximamente ajustado e maximamente folgado em uma projeção de uma região.

O Uso de Sequências de Categorias Gerais Confiáveis para Restringir Percepções

[091] Sistemas típicos de reconhecimento de voz trabalham através de reconhecimento de detalhes, tais como classes de fonemas ou subfonema, e com o uso destes detalhes para determinar padrões de nível mais alto, tais como palavras. Estes detalhes de baixo nível não são distinguidos com certeza, em vez disso são feitas estimativas de probabilidade para cada uma das classes dado um vetor de observação de valores de característica. Modelos Markov Ocultos

(HMM) usam as estimativas de probabilidade de classe juntamente com probabilidades de transição para computar a sequência mais provável de sons de voz pretendidos. Embora a abordagem de "construir a partir dos detalhes" seja popular e razoavelmente eficaz, a mesma não resulta em sistemas de reconhecimento de voz automáticos que rivalizem com o desempenho humano. Um dos inconvenientes desta abordagem é o fato de que classificações detalhadas não são muito confiáveis e precisam ser corrigidas aplicando níveis de contexto mais altos. Além disso, classificações detalhadas são altamente dependentes de contexto, mas o contexto não é conhecido quando se determina a identidade das classes de voz. Adicionalmente, o contexto pode ser representado de forma imprecisa ou com baixa confiabilidade. Além disso, estatísticas precisas são difíceis de estimar para detalhes em contextos que ocorrem raramente. Variações de condições acústicas ou na maneira de falar que não são representadas nas distribuições estatísticas do modelo fazem com que as estimativas estatísticas se tornem imprecisas. Finalmente, o grande espaço de pesquisa de soluções alternativas pode ser intratável computacionalmente. Tipicamente a pesquisa é reduzida por meios arbitrários tal como reter apenas os "n" mais prováveis. Os objetivos da presente invenção são superar os problemas e limitações inerentes na abordagem comum.

[092] Em geral, a classificação em categorias amplas pode ser executada com mais segurança do que a classificação em categorias detalhadas. Por exemplo, distinguir entre um peixe e um pássaro pode ser feito com mais segurança do que determinar os tipos específicos de pássaros ou peixes. Igualmente, no caso de reconhecimento de voz a categorização ampla pode ser executada com mais segurança do que categorização detalhada.

[093] Adicionalmente, a percepção humana parece operar princi-

palmente em categorizações amplas e considerar detalhes apenas quando existe uma razão para focalizar nos mesmos. Em voz contínua fluente, as palavras apenas raramente são produzidas como o dicionário fala que as mesmas deveriam ser, mas isto causa poucos problemas para os ouvintes humanos desde que estejam presentes evidências suficientes para suportar uma percepção. Consequentemente, os ouvintes humanos podem tolerar substituições e omissões desde que aspectos da voz fiquem dentro de categorias amplas confiáveis esperadas geralmente seguindo o sincronismo esperado da voz.

[094] Por exemplo, considerar a pergunta e resposta: "Why you cryin?", "See hit me!". A pergunta omitiu a palavra "are" e substituiu a sílaba "in" por "ing". Nenhuma destas mudanças tem muito efeito na percepção humana. Igualmente, a resposta deve ser mais provavelmente percebida como "She hit me!" mesmo considerando que o som "sh" requerido tenha se tornado similar ao som de "s". A substituição e omissão de detalhes nestes exemplos têm pequeno efeito na percepção e provavelmente devem passar despercebidas por um ser humano. Parece que os padrões de sequências de categorias amplas de sílabas são suficientes para indexar unidades perceptivas que em muitos casos levam a uma percepção ambígua sem exigir identificação específica de classes detalhadas.

[095] A invenção é baseada nas seguintes observações:

- Em grande parte, o padrão de sequência de categorias amplas de voz pode limitar as alternativas perceptivas possíveis. As alternativas perceptivas formam um agrupamento perceptivo.
- A própria sequência de categorias de voz amplas pode ser usada para acessar diretamente a lista de alternativas perceptivas.
- Esforço computacional adicional é aplicado apenas quando necessário para diferenciar entre as alternativas remanescentes dentro de um agrupamento perceptivo.

- Devido às alternativas no agrupamento serem conhecidas em tempo de treinamento, para cada agrupamento perceptivo o processo de diferenciação pode ser otimizado para segurança máxima ou esforço computacional mínimo. Consequentemente, as distinções mais seguras em qualquer circunstância podem ser aplicadas. Isto significa que informação de várias fontes pode ser aplicada, incluindo estatísticas de palavra, prosódicos, gramática, etc.

- Quando diferenciando entre percepções alternadas, a fonética e contextos da palavra das alternativas são conhecidos, deste modo limitando as computações para distinguir características para aquelas que são relevantes e mais seguras. Além disso, detectores e classificadores específicos de contexto podem ser usados para maior segurança.

[096] De acordo com estas modalidades, é apenas quando os padrões de sequência de categorias amplas não diferenciam completamente a percepção, que é exigido recorrer a detalhe. Mesmo neste caso é possível usar preferencialmente discriminações detalhadas que sejam conhecidas como mais seguras do que outras discriminações detalhadas. Por exemplo, considerar um padrão de sequência de categorias amplas de sílaba indexadas a duas percepções que foram distinguíveis uma da outra por diferentes fonemas em duas localizações. Se um dos pares de fonemas era reconhecido como sendo mais seguramente distinguido do que o outro, a distinção deve ser feita na classificação mais segura.

[097] Igualmente, o contexto é muito importante para a percepção. Se a resposta no exemplo dado anteriormente tivesse sido "cuz see hit me!", poderia ser percebida como "cause, he hit me!". Os detalhes do segmento "see" não mudaram, mas a percepção não depende dos detalhes daquele segmento.

[098] Em algumas modalidades da invenção, um algoritmo único

é usado para classificar voz em agrupamentos perceptivos e diferenciar entre percepções alternativas através de acessar otimamente informação disponível. De acordo com estas modalidades, em cada etapa de tempo (ou seja, chegada de outro padrão de sílaba ou sílaba nula se não ocorre nenhuma voz dentro de um certo tempo), o algoritmo classifica a voz dentro de uma sequência de padrões amplos porém confiáveis, tais como categorias amplas de sílabas. Em seguida, cada categoria ampla é associada com um número de categoria. Preferencialmente, categorias similares têm designados números similares.

[099] Em seguida, o algoritmo mapeia sequências de categorias amplas em padrões de percepção usando os números de categoria como coordenadas no espaço de estado. Cada ponto no espaço de estado é associado com um agrupamento perceptivo e uma estratégia de diferenciação. A estratégia de diferenciação, estabelecida durante o treinamento, é uma sequência de etapas a ser executada quando o agrupamento perceptivo é acessado. O propósito da estratégia de diferenciação é diferenciar entre percepções alternativas para acessar informação disponível otimamente. A estratégia de diferenciação é determinada através de avaliação de exigências computacionais e sucesso de várias técnicas de diferenciação aplicadas em diferentes ordens e em diferentes combinações. O resultado final de aplicar a estratégia é uma redução de percepções alternativas para uma pequena quantidade, preferencialmente uma.

[0100] Se as alternativas são reduzidas a uma única percepção, a percepção é ativada. Em um sistema de voz para texto isto envolve fornecer as palavras que correspondem à percepção. Em um sistema controlado por voz, as ações associadas com a percepção devem ser executadas.

[0101] Se as alternativas não são reduzidas a uma única percepção e um limite máximo de latências foi atingido, a percepção mais

provável é aceita como a percepção e ações são geradas de acordo. Se o limite máximo de latência não foi alcançado as percepções alternativas restantes são retidas e interagem com etapas subsequentes no tempo tanto para ajudar na diferenciação de percepções nestas etapas como para serem diferenciadas através da informação disponível nestas etapas no tempo.

Mecanismo de Reconhecimento de Voz Automático

[0102] Nas modalidades atualmente preferenciais da invenção, é fornecido um aparelho para executar todos os aspectos originais da invenção. Nas modalidades atualmente preferenciais da invenção, o sistema de reconhecimento de voz automático é usado em criação de legendas (closed captioning) de televisão em tempo real e ambientes de detecção de palavra.

[0103] A figura 8A ilustra uma representação de um sistema automático de voz para texto 800 que compreende extração baseada em evento e reconhecimento em uma escala de sílaba de classificações amplas de sílaba. O sistema automático de texto para voz 800 usa padrões de sequências de classificações amplas de sílaba para indexar em listas de unidades perceptivas com referencia ao nível de detalhe de fonema apenas quando necessário para diferenciação. Nas modalidades atualmente preferenciais da invenção, o sistema automático de texto para voz 800 escolhe quais classificações de fonema fazer ou escolhe outros métodos de diferenciação para empregar baseado na segurança destas classificações ou métodos.

[0104] O sistema automático de texto para voz 800 inclui um analisador acústico 802. O analisador acústico recebe um sinal de voz de entrada 801 e digitaliza o dito sinal de entrada 801. O analisador acústico 802 é acoplado opcionalmente com um analisador prosódico 803 e com um extrator de evento 804. Em algumas modalidades da invenção, o sinal digitalizado é processado pelo analisador prosódico 803, deste

modo extraindo várias características linguísticas do orador que incluem, mas não estão limitadas a ritmo, tensão, entonação, ou outra informação prosódica que reflete: o estado emocional do orador; se a expressão é uma afirmativa, pergunta ou comando; ironia; sarcasmo; ênfase; foco; etc. De acordo com estas modalidades, a informação prosódica e o sinal digitalizado são enviados para o extrator de evento 804.

[0105] O extrator de evento 804 compreende um mecanismo de processamento para identificar automaticamente regiões em uma pluralidade de sinais de voz que contém padrões de evento e extrair os ditos eventos para reconhecimento de voz. Nas modalidades atualmente preferenciais da invenção, os processos e métodos descritos acima para reconhecimento e extração de evento são empregados pelo extrator de evento 804. O extrator de evento 804 é acoplado a uma memória de evento de curto prazo 805 para armazenar os eventos de voz extraídos. A memória de evento de curto prazo 805 é acoplada com uma pluralidade de módulos de processamento de fluxo de evento para texto para usar os eventos extraídos para fornecer um fluxo de texto resultante. Nas modalidades atualmente preferenciais da invenção, os módulos de processamento de fluxo de evento para texto compreendem um detector de núcleo de sílaba 806, um categorizador de sílaba 807, um módulo de indexação perceptiva de sequência de sílaba 808, e um módulo de categorização de detalhe de subsílaba 809. Os módulos de processamento de fluxo de evento para texto fornecem um fluxo de texto com informação prosódica adicionada 811 embutida no mesmo.

[0106] O sistema automático de texto para voz 800 mostrado na figura 8A compreende um exemplo de um aparelho para reconhecimento de voz automático e para melhorar o mesmo. Ficará prontamente aparente para os versados na técnica tendo o benefício desta descrição que qualquer quantidade de sistemas, configurações, com-

ponentes de hardware, etc. pode ser usada para executar estes métodos e processos para reconhecimento de voz automático e para melhorar o mesmo.

[0107] A figura 8B ilustra uma representação de um sistema de voz para texto automático 820 que compreende um mecanismo de reconhecimento de voz 824 para processar um sinal de voz de entrada 821 de acordo com algumas modalidades da invenção. Nas modalidades atualmente preferenciais da invenção, um analisador acústico 822 recebe o sinal de voz de entrada 821 e digitaliza o dito sinal de voz de entrada 821. O analisador acústico 822 é acoplado com o analisador prosódico 823 e com um mecanismo de reconhecimento de voz 824. Em algumas modalidades da invenção, o sinal digitalizado é processado através do analisador prosódico 823, deste modo extraindo a informação prosódica, como explicado acima.

[0108] Nas modalidades atualmente preferenciais da invenção, o mecanismo de reconhecimento de voz 824 compreende uma pluralidade de módulos de processamento para executar várias etapas do processamento de reconhecimento de voz. Como mostrado, o mecanismo de reconhecimento de voz 824 compreende: um extrator de evento 825; um identificador de padrão 826; um filtro de região átona 827; um simplificador de conjunto reforçado 828; um identificador de sequência de evento 829; um detector de indicador alternativo 830; um criador de conjunto de detectores de cascata 831; um generalizador de voz 832; e um módulo de diferenciação de agrupamento perceptivo 833. Embora módulos de processamento específicos sejam listados neste documento, ficará prontamente aparente para os versados na técnica tendo o benefício desta descrição que quaisquer ferramentas de reconhecimento de voz, conhecidas atualmente ou desenvolvidas posteriormente, podem ser incorporadas como um módulo de processamento no mecanismo de reconhecimento de voz 824.

[0109] Em algumas modalidades da invenção, o extrator de evento 825 compreende um módulo de reconhecimento de voz baseado em evento para construir um esquema de classificadores ponderado para uso em mecanismo de reconhecimento de voz 824. Em algumas modalidades da invenção, o identificador de padrão 826 identifica automaticamente regiões em uma pluralidade de sinais de voz que contém padrões de evento. Em algumas modalidades da invenção, o filtro de região átona 827 empregou diversas técnicas para filtrar regiões que são improváveis de resultar em detectores de átona robustos. Em algumas modalidades da invenção, o simplificador de conjunto reforçado 828 reduz a complexidade dos conjuntos de detector criados através de algoritmos de reforço adaptativos. Em algumas modalidades da invenção, o identificador de evento 829 detecta sequências de eventos em vez de, ou adicionalmente a, detecção de eventos individuais. Em algumas modalidades da invenção, o detector de indicador alternativo 830 reconhece indicadores de voz alternativos quando aspectos de sinal de voz são corrompidos. Em algumas modalidades da invenção, o criador de conjunto de detector de cascata 831 cria automaticamente conjuntos de detectores. Em algumas modalidades da invenção, o generalizador de voz 832 melhora a generalização através da maximização da margem geométrica, como explicado acima. Em algumas modalidades da invenção, o módulo de diferenciação de agrupamento perceptivo 833 diferencia voz usando agrupamento perceptivo, como explicado acima. De acordo com estas modalidades da invenção, o mecanismo de reconhecimento de voz 824 fornece dados de voz.

[0110] Em algumas modalidades da invenção, os dados de voz reconhecidos são armazenados em um ou mais bancos de dados 834, em que o um ou mais bancos de dados 834 é preferencialmente acoplado com uma rede 835. Em algumas outras modalidades da invenção, os dados de voz reconhecidos são automaticamente enviados

para uma memória de evento de curto prazo 836 para processamento de voz para texto.

[0111] Em algumas modalidades da invenção, a memória de evento de curto prazo 836 é acoplada com a pluralidade de módulos de processamento de evento para texto para usar os eventos extraídos para fornecer um fluxo de texto resultante. Nas modalidades atualmente preferenciais da invenção, os módulos de processamento de evento para texto compreendem um detector de núcleo de sílaba 837, um categorizador de sílabas 838, um módulo de indexação perceptiva de sequência de sílabas 839, e um módulo de categorização de detalhe de subsílaba 840. Os módulos de processamento de fluxo de evento para texto fornecem um fluxo de texto com informação prosódica adicionada 841 embutida no mesmo.

[0112] Em algumas outras modalidades da invenção, é fornecido um aparelho para extrair dados de eventos a partir de um sinal de voz e detecção de palavra no mesmo. A figura 8C ilustra uma representação de um sistema 850 para reconhecimento de evento e detecção de palavra que compreende extração baseada em evento e reconhecimento de palavras específicas. O sistema de voz para texto automático 850 inclui um analisador acústico 852 para receber um sinal de voz de entrada 851. O analisador acústico 852 é opcionalmente acoplado com um analisador prosódico 853 e com um extrator de evento 854. O extrator de evento 854 compreende um mecanismo de processamento para identificar automaticamente regiões em uma pluralidade de sinais de voz que contém padrões de evento e extrair os ditos eventos para detecção de palavra. O extrator de evento 854 é acoplado com uma memória de evento de curto prazo 855 para armazenar os eventos de voz extraídos. A memória de evento de curto prazo 855 é acoplada com uma pluralidade de módulos de processamento de detecção de palavra. Em algumas modalidades da invenção, os módulos de pro-

cessamento de detecção de palavra compreendem um detector de núcleo de sílaba 856 e um detector de palavra 857. Os módulos de processamento de detecção de palavra iniciam uma ou mais ações quando uma palavra é detectada.

[0113] O segundo módulo de processamento 862 compreende um classificador de rede neural pulsada. A informação usada para percepção de voz não é uniformemente distribuída em frequência, amplitude e tempo. Padrões temporais são muito importantes para reconhecimento de voz. As redes neurais pulsadas permitem codificação da informação da voz em padrões temporais de pulsos e as estruturas de memória indistintas permitem tolerância de variabilidade temporal. O terceiro módulo de processamento 863 compreende um ou mais mecanismos de reconhecimento de voz consecutivos, como explicado abaixo.

[0114] O sistema de voz para texto alternativo 860 também inclui um analisador acústico 866 para analisar e digitalizar sinais de voz de entrada 867. Os sinais de voz digitalizados são processados por um ou mais dos três módulos de processamento 861, 862 ou 863 e os resultados são alimentados para um módulo de decisão 868, que escolhe os resultados melhor reconhecidos e entrega uma saída de texto 869.

[0115] Algumas modalidades da invenção envolvem segmentação de um sinal de voz em localizações importantes perceptivamente. Isto proporciona um meio para extrair não apenas sincronizações relevantes perceptivamente, mas também sincronizar a análise do sinal com eventos de voz, deste modo evitando todos os problemas de análise de quadro fixo assíncrono, como discutido acima.

[0116] O método primeiro executa um filtro de pré-segmentação usando filtros de baixa complexidade que são baseados em certos aspectos da percepção humana e no nos fenômenos de voz que os mesmos são projetados para detectar. Estes filtros detectam as locali-

zações dos padrões perceptíveis indicativos de começo, término, rajadas, pulsos glotais, e outros eventos de sinal de voz significativos.

[0117] A filtragem de evento pré-segmentação define intervalos que são usados para sincronizar certas computações de categorias. Os padrões de características que tiverem sido extraídos sincronamente são adicionalmente processados para criar características sobre escalas de tempo mais longas e detectar níveis ainda mais altos de eventos perceptivos tais como limites de fonema, núcleo de sílaba, etc.

[0118] A figura 9 ilustra um exemplo de segmentações de um sinal de voz de acordo com algumas modalidades da invenção. O sinal de voz da figura 9 contém a expressão "Once". O sinal muda de feitio diversas vezes ao longo do curso da expressão de formas que são visualmente aparentes quando visualizando a forma da onda. As segmentações indicadas pelas marcas verticais curtas no fundo do gráfico correspondem a eventos de pulso glotal durante a parte "sonora" da palavra.

[0119] As linhas verticais longas correspondem a vários tipos de eventos de limite de som de voz. Para referência, os rótulos dos segmentos foram colocados no gráfico que indica a identidade fonética do segmento. As condições do sinal nas transições entre fonemas variam pelo tipo de transição. Em alguns limites a energia total muda abruptamente, enquanto para outros mudanças espectrais são associadas com o evento. Tomados em conjunto, estes vários eventos permitem que extração de característica seja executada sincronamente com os eventos de voz e fornecem segmentação relevante perceptivamente.

[0120] Em algumas modalidades da invenção, a segmentação de sinal é baseada em diferenças perceptivas presentes no sinal de voz. Frequentemente, a informação usada para percepção de voz não é distribuída uniformemente no tempo. A percepção humana é sensível a mudanças em estímulos. Em sinais temporais tais como voz, as lo-

calizações de tempo de mudanças significativas (ou seja, eventos), proporcionam a organização perceptiva do sinal. O sincronismo relativo dos eventos e as características dos estímulos em sua vizinhança codificam muito da informação perceptiva. Em geral, as percepções de magnitude não são lineares. Por exemplo, é sabido que a percepção da intensidade de som é logarítmica e medida comumente em decibéis. Pode ser demonstrado que, para uma ampla gama de percepções, a diferença apenas perceptível no estímulo é relacionada ao nível original do estímulo. Entretanto, isto não se sustenta nos extremos e não existe percepção na extremidade inferior até que o nível de estímulo alcance um nível mínimo para ativação neural. Na extremidade superior, uma vez que os neurônios comecem a saturar, aumentos adicionais no estímulo não são percebidos. Na amplitude operacional, para muitos tipos de estímulo, a mudança necessária para uma resposta perceptiva pode ser aproximada pela lei de Weber: $K = \Delta I / I_0$; em que $I_0$ é o nível de estímulo original, $\Delta I$ é a mudança no nível de estímulo, e K é constante determinada empiricamente que define o limite da diferença apenas perceptível.

[0121] O lado direito da formulação da lei de Weber pode ser reconhecido como contraste. Na presente invenção, eventos são declarados (ou seja, o detector dispara) quando a mudança em uma característica relevante excede um limite perceptível. Na presente invenção, a mudança perceptiva é computada com o uso de um cálculo de contraste perceptivo relacionado à lei de Weber.

[0122] A figura 10 ilustra uma fórmula de contraste perceptivo usada para computar mudança perceptiva de acordo com algumas modalidades da invenção. Nesta fórmula, o denominador da relação do lado direito difere da formulação padrão da lei de Weber de duas formas: a mesma inclui a soma dos valores que são contrastados e a mesma inclui um fator adicional ε. O fator ε inibe a ativação em níveis

muito baixos para simular melhor a resposta perceptiva para estímulos de muito baixo nível. O mesmo também torna a fórmula numericamente estável evitando uma divisão por zero quando nenhum estímulo está presente.

[0123] A inclusão da soma dos valores contrastantes nivela adicionalmente a resposta de contraste perceptivo em níveis muito baixo e muito alto. Para cada característica perceptiva medida (por exemplo, energia ou frequência), valores apropriados de ε e limites perceptivos são estabelecidos empiricamente. Em algumas modalidades da invenção, é criada uma pluralidade de detectores de eventos perceptivos heterogêneos, em que cada um é baseado em alguma característica de sinal particular, medida em alguma escala de tempo particular, e com seu ε e limites perceptivos particulares.

[0124] Os detectores de eventos da invenção operam em vários aspectos do sinal em várias escalas. Primeiro, a pré-segmentação é executada através do processamento de valores de energia através de filtros de baixa complexidade que detectam as localizações temporais das rajadas, fechamentos e pulsos glotais. A extração de características então é executada relativa aos eventos de pré-segmentação. Filtros e detectores adicionais são aplicados às características extraídas sincronamente para extrair características e eventos de alto nível.

Técnicas de Processamento e Extração de Característica Adicional

Memória de Fila Circular Segmentada

[0125] Diversos componentes de detectores de evento envolvem comparações de somas de valores de características computadas com o uso de janela de análise de várias extensões, alinhadas em relacionamentos temporais específicos uma com respeito à outra. Para minimizar a carga computacional dos detectores de evento estas somas são mantidas com o uso de uma memória de fila circular segmentada. Uma fila circular é uma estrutura de memória do tipo, primeiro a entrar

primeiro a sair (FIFO) onde nova informação é gravada na memória em $I_0$, o índice da informação mais antiga na memória. Após gravar a nova informação na memória o índice $I_0$ é avançado um módulo do tamanho da memória (ou seja, o índice $I_0$ volta para zero quando chega ao fim da memória). As somas correntes dos valores na memória podem ser mantidas de acordo com o processo descrito abaixo.

[0126] Primeiro, inicializam-se as localizações de memória de fila circular, a soma corrente, e o índice $I_0$ para zero. Em seguida, em cada etapa: subtrai-se o valor indexado a partir da soma corrente; adiciona-se o novo valor a soma corrente; grava-se o novo valor dentro da fila circular e avança-se o índice $I_0$ um módulo da dimensão da memória.

[0127] A operação de uma fila circular e sua utilidade para a comutação eficiente de somas correntes é ilustrada nas figuras 11A a 11C. A figura 11A ilustra uma memória de fila circular de acordo com algumas modalidades da invenção. Na figura 11A, uma memória de fila circular de 5 elementos é representada no momento "t" quando um novo valor, "7", está para ser gravado. O novo valor irá sobrescrever o mais velho na memória que, no exemplo ilustrado, tem um valor de "9". Antes de gravar o novo valor, a soma dos valores na memória de exemplo é 25. Devido ao novo valor sobrescrever o valor mais antigo, a soma corrente pode ser mantida através da subtração do valor antigo e soma do valor novo. Como pode ser visto prontamente, a complexidade computacional de manter somas correntes desta maneira é independente da dimensão da memória. Apenas uma subtração e uma adição são requeridas independente do tamanho da memória.

[0128] A figura 11B e a figura 11C ilustram uma memória de fila circular atualizada de acordo com algumas modalidades da invenção. Mais especificamente, a figura 11B e a figura 11C mostram o processo de atualização continuando através das duas próximas etapas. Para manter somas correntes de valores sobre várias subseções de memó-

ria, a fila circular é segmentada através do uso de índices adicionais, cada um dos quais tem um deslocamento fixo a partir do índice $I_0$. Cada uma das subseções de soma corrente é mantida simplesmente pela subtração do valor que está próximo a ser retirado da subseção e adição do valor que está próximo a se tornar parte da subseção.

[0129] A figura 12 ilustra uma fila circular segmentada para manter duas somas correntes de acordo com algumas modalidades da invenção. A fila circular segmentada é disposta para facilitar a manutenção de duas somas correntes, uma computada para a metade mais antiga dos valores na fila circular (ou seja, subseção A) e a outra computada para a metade mais recente dos valores na fila circular (ou seja, subseção B). Estas somas são referenciadas como $\Sigma_A$ e $\Sigma_B$ respectivamente. Agora existe um segundo índice $I_1$ mantido em um deslocamento igual a uma metade da dimensão da memória a partir do índice $I_0$. Em cada etapa no tempo o valor indexado por $I_0$ (ou seja, o valor mais antigo em toda a memória) é subtraído de $\Sigma_A$ e o valor indexado por $I_1$ é somado a $\Sigma_A$, enquanto o valor indexado por $I_1$ é subtraído de $\Sigma_B$ e o novo valor a ser escrito na memória é somado a $\Sigma_B$. O novo valor é gravado na localização no índice $I_0$, e os índices $I_0$ e $I_1$ são em seguida incrementados um módulo do tamanho da memória. No exemplo agora apresentado, as subseções da memória são de dimensão igual, formam conjuntos desmembrados, e juntas cobrem a memória inteira. Nenhuma destas condições é exigida pelo método.

[0130] A figura 13 ilustra uma fila circular segmentada de acordo com algumas modalidades da invenção. Na figura 13, a subseção "A" é disposta de modo que a mesma fica completamente dentro da subseção "B". A dimensão total da memória bem como as dimensões de cada subseção e a disposição temporal de subseções são determinadas de acordo com o propósito para o qual as somas estão sendo mantidas.

[0131] Em algumas modalidades da invenção, as filas circulares são usadas para detectar localizações de mudanças abruptas. Diversos eventos de voz importantes, tais como começo, término, paradas de rajadas, etc., são associados com mudanças semi-monotônicas abruptas nos níveis de algumas características do sinal. Uma fila circular segmentada disposta em geral como na figura 13 pode ser empregada para detectar mudanças semi-monotônicas abruptas. Com as dimensões de subseções "A" e "B" determinadas adequadamente, a diferença perceptiva entre somas correntes das subseções "A" e "B" é computada a cada etapa de tempo. Os tempos onde a diferença perceptiva alcança um máximo e sua magnitude excede seu limite perceptivo se tornam candidatos a pontos de segmentação. Qualificações adicionais são aplicadas às características que imitam mais proximamente a percepção humana forçando uma separação de tempo mínima entre eventos detectados. Já neste estágio, os eventos podem começar a ser classificados grosseiramente, baseados na direção da mudança no evento. Por exemplo, eventos devido a fechamentos são diferenciados de começos e rajadas pela direção da mudança de energia através da transição.

[0132] Em algumas outras modalidades da invenção, as filas circulares são usadas na detecção de impulsos e lacunas nos sinais de voz. Alguns eventos de voz importantes são associados com localizações no tempo onde algumas características do sinal mudam abruptamente por um período muito breve de tempo e em seguida retornam a um nível similar àquele que estavam antes da mudança. Se a mudança breve é para um valor maior a mudança é chamada um "impulso". Se a mudança breve é para um valor menor a mudança é chamada uma "lacuna". Uma fila circular segmentada disposta em geral como na figura 5 pode ser empregada para detectar impulsos e/ou lacunas. Com as dimensões das subseções "A" e "B" determinadas adequadamente, impulsos

(lacunas) são localizados quando o valor médio na subseção "A" está acima (abaixo) do valor médio na subseção "B" por um valor limite adaptativo perceptível. Como explicado previamente, a função de limite é determinada empiricamente. Os tamanhos das subseções "A" e "B" são determinados de acordo com a natureza da percepção humana e características temporais dos aspectos dos sinais a serem detectados.

Detecção de Pulso Glotal

[0133] Um caso especial importante que ilustra o uso desta abordagem é a detecção de eventos de pulso glotal. Eventos de pulso glotal são localizados através do seguinte procedimento. Primeiro, o sinal é filtrado em passa banda na amplitude do primeiro formante. Em seguida, a energia Teager é computada como Teager (t) = x(t) * x(t) – x(t-1) * x(t+1); em que x(t) é o valor de entrada no tempo t.

[0134] Sendo uma função da amplitude e frequência, a energia Teager enfatiza as localizações de pulso dos pulsos glotais, as quais são associadas com máximos locais dos componentes de energia e alta frequência. Finalmente, o sinal é segmentado com o uso de um detector de impulso disposto em geral como na figura 13. O detector é baseado em somas correntes de valores absolutos da energia Teager. Na modalidade preferencial, as dimensões de subseções "A" e "B" são ajustadas para 2 ms e 10 ms respectivamente. O detector está em um estado alto sempre que a energia Teager média na subseção "A" é maior do que o limite perceptivo K multiplicado pela energia Teager média na subseção "B". O valor de K foi escolhido para ser 1,3. As dimensões das subseções "A" e "B", e o valor do multiplicador K têm sido considerados úteis para detectar localizações de pulso glotal. Valores diferentes destes descritos aqui podem ser usados dentro do escopo desta invenção.

[0135] O detector de pulso glotal há pouco descrito cria dois eventos de localizações para cada pulso glotal, um no limite ascendente do pulso e um no limite descendente do pulso. O período de afastamento

é definido como o período entre dois eventos de limites ascendentes sequenciais. A duração do pulso é estimada pelo tempo entre o limite ascendente e o limite descendente subsequente. A relação da duração do pulso para o período de afastamento total é relacionada ao "quociente aberto", uma característica da voz sonora que pode ser útil em algumas aplicações de processamento de voz. Além disso, durante a parte aberta do período do afastamento as cavidades subglotais são acusticamente acopladas com as cavidades orais criando padrões de formante um pouco diferentes durante esta parte comparada aos padrões da parte fechada. Este fato pode ser explorado vantajosamente dispondo a extração de característica em relação a estes eventos.

[0136] A figura 14 ilustra uma representação de uma saída do detector de pulso glotal em um segmento de voz sonora de acordo com algumas modalidades da invenção. Na figura 14, a saída do detector de pulso glotal divide o sinal em "segmentos "alto" e "baixo". Os segmentos representam tempos em que uma característica relevante (neste caso energia Teager) está perceptivamente acima da norma. Esta disposição cria um segmento para a duração do pulso ou lacuna. Para algumas aplicações pode ser preferível marcar um pulso ou lacuna em vez de um segmento. Nestes casos a seleção de tempos de evento específicos pode ser determinada por um dos diversos métodos alternativos que incluem, mas não limitados a:

- selecionar o ponto médio entre os limites ascendente (descendente) e descendente (ascendente);
- selecionar o limite ascendente do segmento;
- selecionar o limite descendente do segmento;
- selecionar o valor de característica máximo (mínimo) dentro do segmento; e
- selecionar o ponto do contraste perceptivo máximo dentro do segmento.

[0137] Detecção de pulso glotal como delineado acima é baseada na detecção quando o valor médio de uma certa característica do sinal (por exemplo, energia Teager) dentro de uma janela disposta centralmente desvia significativamente da média da mesma característica calculada por um período de tempo mais longo. Filas circulares segmentadas dispostas em geral como na figura 13 podem ser usadas para segmentar qualquer sinal de modulação identificando as regiões onde uma característica de voz selecionada (por exemplo, energia ou frequência formante) desvia perceptivamente de sua norma de longo prazo. Devido a o custo computacional para manter as somas correntes usadas pelos detectores ser independente da dimensão das subseções, as mesmas podem ser usadas para segmentar modulações de grande escala tão bem quanto impulsos breves.

Detecção de Núcleo de Sílaba

[0138] Para ilustrar este ponto, um detector de núcleo de sílaba foi construído usando uma fila circular segmentada disposta em geral como na figura 13, para manter somas correntes da energia de Teager, computadas exatamente como para o detector de pulso glotal exceto a dimensão da subseção "A" que foi ajustada para 60 ms e a dimensão da subseção "B" que foi ajustada para 100 ms.

[0139] A figura 15 ilustra uma representação de uma forma de onda de saída de acordo com algumas modalidades da invenção. A figura 15 mostra as saídas de forma de onda e detector para a palavra "Once" falada duas vezes, a primeira normalmente e a segunda vez em um sussurro. Como pode ser visto este detector geralmente agrupa as partes centrais das sílabas.

[0140] Algumas modalidades da invenção envolvem métodos para reconhecer padrões de voz usando extração de formante. Conforme a voz é produzida, as configurações dos órgãos de articulação (por exemplo, língua, mandíbula, lábios) criam padrões dinâmicos de res-

sonâncias e antirressonâncias nos espectros de frequência chamados formantes. Durante a voz sonora, o som é gerado tanto pelos "ruídos de ar" difusos como por estrutura harmônica fortemente organizada. Ambos os componentes difuso e harmônico contribuem para o entendimento da voz e ambos são variavelmente invocados sob diferentes condições de ruído. Os "ruídos de ar" difusos interagem com os formantes e são formatados pelos mesmos, revelando-os para ser relativamente atenuados. Os harmônicos resolvidos fortes criam picos relativamente agudos no espectro e, se não processados apropriadamente, fazem com que seja difícil localizar precisamente formantes próximos. As séries de harmônicos fornecem um excelente meio para determinar o afastamento, mesmo quando a própria frequência do período de afastamento está faltando no sinal. Experimentos têm mostrado que os harmônicos de amplitude modulada podem ser usados para recriar voz inteligível que "ignora" ruído. Durante voz não sonora mudanças perceptíveis temporalmente dividem o sinal em segmentos semi-homogêneos.

Extração de Formante

[0141] Em algumas modalidades da invenção, um processo de extração de formante é executado como descrito na figura 16. A figura 16 ilustra um fluxo de trabalho 1600 para executar extração de formante de acordo com algumas modalidades da invenção.

[0142] O fluxo de trabalho 1600 começa quando as amostras do segmento são ajustadas à janela Hamming 1601 com uma extensão de janela igual à extensão do segmento, em que o segmento corresponde a um período de afastamento durante a voz sonora. As amostras ajustadas a janela são em seguida processadas através de um banco de filtros passa banda larga 1602. Em algumas modalidades, os filtros passa banda têm larguras de banda de 400 Hz e são espaçados em centros de 50 Hz cobrindo a amplitude de 450 Hz até 4.000 Hz.

Em seguida, o fluxo de trabalho computa a amplitude instantânea e a frequência de cada filtro é computada com o uso da técnica DESA-1 16 03. Baseadoa em suas qualidades numéricas, os valores computados são julgados como "válidos" ou "não válidos" na etapa 1604. Em seguida, conta e armazena estimativas "válidas" em uma memória temporária.

[0143] Em seguida, um histograma cujas caixas representam amplitudes de frequência é inicializado 1606, em que para cada estimativa válida, a caixa de histograma que representa a frequência instantânea estimada é incrementada pelo logaritmo da amplitude instantânea estimada comprimida correspondente. Em seguida, os picos do histograma atenuado são selecionados como candidatos a formante 1607, as frequências, larguras de banda (sigmas) e amplitudes de formante são retidas como características 1608, e as características delta são computadas nas trilhas de formante por ajustamento linear 1609. Finalmente, em localizações de mudança perceptível nos padrões de formante, são gerados eventos 1610.

Processamento de Banco de Filtros de 12ª Oitava

[0144] Em algumas outras modalidades da invenção, é executado um processo de processamento de banco de filtros de 12ª oitava no sinal segmentado usando passa bandas estreitos nas frequências mais baixas e passa bandas mais largos nas frequências mais altas imitando as tendências de resolução de frequência encontrados na audição humana. A figura 17 lustra um fluxo de trabalho 1700 para executar extração de formante de acordo com algumas modalidades da invenção.

[0145] O fluxo de trabalho 1700 começa quando as amostras do segmento estão síncronas com o sinal, ajustado à janela Hamming 1701 com uma extensão de janela igual à extensão do segmento, em que o segmento corresponde a um período de um afastamento. Em

seguida, as amostras ajustadas à janela são processadas através da um banco de filtros espaçados de 12ª oitava 1702 e a amplitude e frequência instantâneas de cada filtro são computadas com o uso da técnica DESA-1 1703. Baseados em suas qualidades numéricas, os valores conjugados são julgados 1704 "válido" ou "não válido", em que estimativas "válidas" são contadas e armazenadas em uma memória temporária para o intervalo 1705.

[0146] Em seguida, é construído um histograma, cujas caixas correspondem às frequências centrais de cada filtro no banco de filtros de 12ª oitava 1706, em que para cada valor estimado, a caixa do histograma cuja amplitude inclui frequência instantânea estimada é incrementada pelo logaritmo da amplitude instantânea estimada comprimida. Em seguida, os pesos do histograma são multiplicados por uma função de ponderação baseada na sensitividade do ouvido em diferentes frequências 1707. Após computar os histogramas, os padrões de energia de caixa do histograma são somados em combinações harmônicas para detectar a sequência harmônica mais forte com a energia mais forte 1708, em que o fundamental da sequência harmônica é usado como uma estimativa de afastamento. Se a aplicação requer estimativas ainda mais precisas, filtros passa banda estreitos são centrados nas frequências harmônicas estimadas e recomputados 1709. Este processo converge rapidamente em estimativas altamente precisas. Finalmente, a relação de energia harmônica para a energia total é computado como uma medida de sonoridade 1710, em que os padrões de relação de amplitude dos harmônicos são mantidos como características, em que a relação é usada em reconhecimento de voz automático.

Uso dos Períodos de Afastamento

[0147] Em algumas modalidades da invenção, os começos e afastamentos das trilhas de harmônicos podem ser determinados por am-

plitudes relativas de período de afastamento para período de afastamento. Mudanças abruptas na amplitude das trilhas de harmônicos são associadas com a interação dos harmônicos com os formantes, e as mudanças abruptas indicam uma mudança na interação, que pode ser devida a uma mudança no afastamento ou uma mudança no formante. Estas mudanças são indicativas de uma localização de transição. Eventos podem ser gerados em resposta a estas mudanças com o uso dos métodos de filtro expostos previamente. Deve ser observado que estes eventos, quando ocorrem, serão síncronos com o sincronismo dos pulsos glotais.

Normalização do Trato Vocal e Reconhecimento de Segmento de Fonema Atenuado

[0148] Em algumas modalidades da invenção um processo de normalização do trato vocal e reconhecimento de segmento de fonema atenuado é empregado para solucionar complicações inerentes ao uso de padrões de formantes como características. Os padrões de formantes gerados por um orador codificam simultaneamente informação sobre os sons da voz que são produzidos e a extensão do trato vocal do orador. Isto complica o uso dos padrões de formato como características.

[0149] Foi observado em Watanabe, e outros, *Reliable methods for estimating relative vocal tract* lengths *formant trajectories of common words, IEEE* transactions on audio, speech, and language processing*, 2006, vol. 14 pp. 1193 a 1204,* que os formantes para dois oradores produzindo o mesmo som de voz têm um relacionamento inversamente proporcional à relação de suas extensões de tratos vocais:

$$L_A/L_B = F_{nA}/F_{nB}$$

[0150] Conforme sons de voz diferentes são produzidos, a extensão do trato vocal do orador é modificada continuamente através da reconfiguração dinâmica dos órgãos de articulação. Para um dado orador, conforme cada som é produzido os formantes se moverão para

cima ou para baixo porque os mesmos estão modificando a extensão do trato vocal. Aplicar a fórmula de Watanabe ao padrão de formante do orador "A" pronunciando um certo som de voz e o padrão de formante do orador "B" pronunciando o mesmo som, fornece uma estimativa de suas extensões de trato vocal relativas para cada formante medido. Alguns aspectos da invenção são baseados nas informações a seguir. Primeiro se um orador "A" e orador "B" estão produzindo o mesmo som, as estimativas de trato vocal relativas baseadas em cada um dos vários formantes medidos irão aproximar o valor real e, portanto um será similar ao outro. Em seguida, se o orador "A" e orador "B" estão produzindo sons diferentes, as estimativas de extensão de trato vocal baseadas em cada um dos vários formantes medidos será divergente. Adicionalmente, se a transição de um certo som de voz envolve alongamento (encurtamento) da extensão do trato vocal quando falado pelo orador "A", o mesmo também envolverá o alongamento (encurtamento) da extensão do trato vocal do orador "B" mas por quantidades diferentes baseado em sua fisiologia.

[0151] Em algumas modalidades, os valores de formantes para cada som de voz, como falada por um orador de referência, são gravados. As medições de formante do orador de referência podem ser baseadas em um único orador ou mais, preferencialmente ser tomadas como a média de medições de muitos oradores. No momento do reconhecimento, cada segmento é processado para produzir valores de formante como descrito previamente. Cada som de voz (ou seja, fonema ou fonema parcial) é por sua vez assumido como sendo o que está sendo falado, e os valores de formante do segmento atual são usados para computar estimativas de extensão de trato vocal relativo do orador corrente para aquele do orador de referência. Baseado na lista de consistências, a probabilidade relativa de cada som de voz pode ser estabelecida. Conforme a trajetória da voz se aproxima da con-

figuração-alvo de cada padrão de formante aprovado, a consistência das estimativas irá aumentar e nestes tempos alvo tenderão a ser maiores para o som de voz percebido. A confiança que pode ser aplicada a tais percepções é dependente das condições do som de voz e ruído. Quando sons de voz são determinados com alta confiança, os mesmos se tornam pontos de referência no sinal úteis para restringir os possíveis padrões nas regiões com menor confiança.

Mecanismos de Reconhecimento de Voz Automáticos Paralelos Consecutivos

[0152] Algumas modalidades da invenção envolvem usar uma pluralidade de mecanismos de reconhecimento de voz automáticos (ASR) paralelos consecutivos em modo de rajada se sobrepondo temporariamente para reduzir a latência e melhorar a precisão. Cada mecanismo ASR pode ser de concepção e origem similar ou diferente, mas todos têm que ser capazes de produzir resultados aceitáveis na linguagem-alvo na parte central do segmento dentro do quadro de tempo de segmentação mínimo. Os resultados dos processadores consecutivos são analisados pela ponderação das palavras produzidas durante a parte central de cada segmento maior do que as palavras produzidas no início e no fim, sincronização dos segmentos pela melhor adaptação, e as palavras com maior peso são selecionadas para saída.

[0153] Estas modalidades envolvem o uso de múltiplos mecanismos ASR nos segmentos de voz de áudio sobrepostos para reduzir a latência e melhorar a precisão. A abordagem paralela consecutiva aumenta a precisão ao mesmo tempo em que reduz a latência.

[0154] Por exemplo, se um ASR segmenta arbitrariamente um sinal de voz de entrada em x segundos, a saída tende a ser mais precisa na localização x/2, e menos precisa no início e fim do segmento, uma vez que o contexto mais alto em ambas as direções para frente e para trás é encontrado na localização central. Dado este comporta-

mento observado, alguém pode ser capaz de usar esta informação como alavanca simplesmente pela execução de n instâncias de um mecanismo ASR em modo em lote, segmentando o sinal de entrada em rajadas de x segundos que se sobrepõem por x/n segundos, e alternando o encaminhamento destes segmentos entre cada mecanismo. Se n = 2, ao mesmo tempo em que o mecanismo B está trabalhando no reconhecimento do seu segmento, a saída do mecanismo A é analisada juntamente com o fluxo de palavras saído previamente para reforçar estatisticamente, corrigir e fornecer as palavras a partir do mecanismo A. Em seguida, no limite n da segunda entrada, as tarefas de analisador de saída e processamento comutam as obrigações entre os mecanismos.

[0155] Observando um típico mecanismo ASR útil em uma configuração consecutiva, vê-se que x parece trabalhar melhor quando estabelecido por volta de três segundos quando usando um modelo de linguagem Inglês WSJ de três mil palavras. Isto permite a possibilidade de usar o mecanismo, que é projetado e otimizado para trabalhar em expressões longas, para ser adaptado para uso em ambientes onde a baixa latência é necessária.

[0156] Em outras palavras, se x = 3, o primeiro segmento de voz em 0,0 a 3,0 segundos será apresentado para transformação para o mecanismo A. O segmento de 1,5 a 4,5 então será apresentado para mecanismo B, etc.

[0157] A figura 18 ilustra uma representação de dois mecanismos de processamento consecutivos, se sobrepondo no tempo, operando em uma sequência de expressões de acordo com algumas modalidades da invenção. Como mostrado na figura 18, as palavras, "is falling from the sky" são fornecidas pelo mecanismo A, e "done the sky today at" vem do mecanismo B. Empregando métodos estatísticos descontando o peso para cada palavra nas extremidades de cada segmento que leva em con-

ta o fator de confiabilidade para aquelas palavras, pode-se terminar com uma sequência de palavras contínua aparente tal como "is falling from the sky today at" com uma latência fixa de 3 segundos.

[0158] A análise de ponderação e o mecanismo de saída podem incluir um ou mais algoritmos nas seguintes categorias bem como outras para determinar quais palavras serão adicionadas à sequência de saída final. Por exemplo, um algoritmo pode envolver ponderação simples das palavras centrais em um segmento com valores maiores do que as palavras nos limites do segmento, indicadores acústicos e prosódicos ganhos do sinal de voz original, análise estatística das palavras para serem fornecidas para reforçar os pesos da saída mais provável, regras gramaticais para selecionar a saída mais provável, ou outros métodos de aprendizagem de máquina e estatísticos.

Pontuador Automático

[0159] Algumas modalidades da invenção envolvem inserção automática de sinais de pontuação em um texto não pontuado. Um pontuador automático é um sistema que insere sinais de pontuação (pontos, vírgulas, pontos de interrogação, pontos de exclamação, apóstrofos, aspas, parêntesis, elipses, ponto e vírgula e dois pontos) em um texto não pontuado.

[0160] A figura 19 ilustra um sistema de voz para texto 1900 que inclui um pontuador automático de acordo com algumas modalidades da invenção. Em algumas modalidades da invenção, texto não pontuado pode originar um texto 1901, ou linguagem vozda 1902 que é em seguida transcrita para texto por um sistema de reconhecimento de voz automático 1903.

[0161] O texto transcrito ou o texto nativo de 1901 é enviado para o pontuador automático 1905. O pontuador automático 1905 cria um texto que é mais facilmente legível e menos ambíguo devido à colocação correta de sinais de pontuação.

[0162] Em algumas modalidades da invenção, o pontuador automático 1905 é acoplado com um banco de dados 1904 que contém dados de treinamento. O pontuador automático usa um ou mais algoritmos Baysianos que é treinado em uma grande quantidade de textos de treinamento que é pontuado corretamente. Os padrões de pontuação nos dados de treinamento são analisados para criar um conjunto de regras que descrevem os padrões de pontuação no texto.

[0163] Uma vez que o pontuador tenha sido treinado em uma quantidade suficiente de texto, suas regras podem então ser aplicadas a um novo texto para predizer onde devem ser inseridos sinais de pontuação.

[0164] Em algumas modalidades da invenção o pontuador automático 1905 compreende uma pluralidade de módulos de processamento. Como mostrado, o pontuador automático inclui um primeiro processador estatístico 1906, um segundo processador estatístico 1907 e um terceiro processador estatístico 1908.

[0165] Em algumas modalidades, o primeiro processador estatístico 1906 identifica lugares onde a pontuação deve ser inserida baseada em regras estatísticas. Um processo de treinamento é conduzido para desenvolver as regras. O processo de treinamento envolve análise das correlações entre palavras específicas e sinais de pontuação em uma grande quantidade de textos pontuados corretamente. O conjunto de regras é derivado desta análise. O conjunto de regras pode então ser aplicado a um novo texto não pontuado para predizer localizações prováveis para sinais de pontuação. A saída deste processo é uma série de opiniões sobre onde os sinais de pontuação devem ser inseridos.

[0166] Em algumas modalidades, o segundo processador estatístico 1907 treina nas correlações de classes de palavra com sinais de pontuação. Este processo é baseado em um marcador de classe de

palavra que analisa a estrutura das sentenças nos dados de treinamento e atribui um rótulo de classe de palavra para cada palavra. Exemplos de rótulos de classe de palavra são substantivo, verbo, adjetivo, preposição, etc.

[0167] O processo em seguida constrói um conjunto de regras baseado em suas observações de como certas classes de palavra se correlacionam com sinais de pontuação. Em seguida o conjunto de regras pode ser aplicado a um novo texto. A saída deste processo é uma série de opiniões sobre onde devem ser inseridos sinais de pontuação dentro do texto.

[0168] Em algumas modalidades, o terceiro processador estatístico 1908 utiliza ponderação baseada em extensões médias de sentença. O terceiro componente do pontuador estatístico é baseado na quantidade de palavras que tipicamente formam sentenças em um texto particular. Como nos outros processos, ele treina em uma grande quantidade de texto pontuado corretamente. As regras são desenvolvidas baseadas na quantidade de n-gramas que ocorrem nas unidades de texto que são limitadas pela pontuação.

[0169] Em algumas modalidades da invenção, os resultados do primeiro processador estatístico 1906 e do segundo processador estatístico 1907 são dois conjuntos de opiniões de onde a pontuação deve ser inserida em um texto. Os resultados do terceiro processador estatístico 1908 são então usados como um tipo de desempate para resolver situações quando as decisões estão em conflito. Por exemplo, se o primeiro processador estatístico 1906 prediz que um ponto é necessário após a quinta palavra em uma sequência de palavras, e o segundo processador estatístico 1907 prediz que um ponto é necessário após a terceira palavra, os resultados do terceiro processador estatístico 1908 devem ser chamados para tomar a decisão, uma vez que é improvável que ambos estejam corretos, porque seria formada uma sentença de duas palavras.

[0170] Em algumas modalidades, o terceiro processador estatístico 1908 atribui um peso maior para os resultados ou do primeiro processador estatístico 1906 ou do segundo processador estatístico 1907 baseado em seu conhecimento da extensão típica de sentença neste tipo de documento. Se as sentenças no tipo de documento são tipicamente muito curtas, o terceiro processador estatístico 1908 deveria atribuir maior peso para a saída do segundo processador estatístico 1907. Se, por outro lado, as sentenças no tipo de documento são usualmente de cinco palavras ou mais longas, ele dever atribuir maior peso à opinião gerada pelo primeiro processador estatístico 1906.

[0171] Uma vez que a etapa de tomada de decisão está completa, o resultado é passado para um módulo de decisão 1909 que fará a decisão final sobre onde inserir pontuação, em conjunto com informação de um módulo de pontuação baseado em regras 1910 e um módulo de afastamento/pausa 1911.

[0172] Em algumas modalidades, um módulo de pontuação baseado em regras 1910 usa um conjunto de regras sobre estrutura linguística para determinar onde sinais de pontuação devem ser inseridos no texto. O módulo de pontuação baseado em regras 1910 é acoplado com um banco de dados léxico 1916.

[0173] O módulo de pontuação baseado em regras 1910 pode identificar diversas classes funcionais de palavras, incluindo pronomes sujeitos, pronomes objetos, pronomes relativos, modais, conjunções, artigos definidos, datas e certas categorias de verbos. Em algumas modalidades, o banco de dados léxico 1916 inclui informação de classe de palavra.

[0174] Uma vez que o programa tenha identificado um membro de uma das categorias funcionais, ele prossegue para a pesquisa do contexto próximo, examinando em uma janela de texto que consiste no item identificado e duas palavras precedentes e seguintes. Categorias

específicas de palavras ou classe de palavra que ocorrem na janela de contexto indicarão a necessidade de uma vírgula em algum ponto na sequência de palavras. As regras de linguística servem como uma lista de instrução para onde as vírgulas devem ser inseridas. Como um exemplo, quando o programa identifica um pronome sujeito (eu, ele, ela, nós, eles) ele verifica a janela de contexto para a ocorrência de outras categorias. Se, por exemplo, o pronome sujeito é precedido por um advérbio ou um particípio (com certos particípios de verbo esperados) o programa irá predizer que ali deve haver uma vírgula após a palavra que precede a palavra identificada. O pontuador baseado em regras pode processar uma sequência de palavras do texto ou um arquivo de texto preexistente. A saída do pontuador baseado em regras é uma série de opiniões sobre onde devem ser inseridas vírgulas.

[0175] Em algumas modalidades o módulo de afastamento/pausa 1911 é diferente dos outros componentes pelo fato de que sua entrada é um arquivo de áudio que contém voz humana. Os outros componentes operam com texto, embora o texto possa ter sido originado como dados de áudio que foram então transcritos. O módulo de afastamento/pausa 1911 opera na observação de que na voz humana, mudanças de afastamento significativas que acontecem sobre um curto período de tempo e são correlacionadas com um período de silêncio são usualmente indicativos de uma necessidade de pontuação. Por exemplo, se um dado ponto de arquivo de áudio mostra uma queda abrupta no afastamento (30% ou mais) ocorrendo em um curto período de tempo (275 ms), que é um indicador provável de que o orador encontrou o fim de uma sentença.

[0176] A presença de uma pausa seguindo este padrão tende a confirmar que uma localização para um sinal de pontuação foi identificada. O pontuador de afastamento/pausa rastreia afastamento de um arquivo de áudio e sinaliza quando as condições corretas foram alcan-

çadas para indicar pontuação. O pontuador de afastamento/pausa fornece opiniões sobre onde sinais de pontuação devem ser inseridos.

[0177] Em algumas modalidades, o módulo de decisão 1909 recebe as entradas do pontuador automático 1905, pontuador baseado em regras 1910, e módulo de afastamento/pausa 1911. Baseado em características conhecidas do tipo de texto, o módulo de decisão 1909 atribui pesos maiores ou menores para cada um destes resultados para fazer uma determinação final sobre se uma pontuação deve ser inserida ou não em um dado ponto do texto.

## REIVINDICAÇÕES

1. Sistema para reconhecimento de voz (800, 820) que corresponde a um sinal de voz digital (801, 821, 851), o sistema (800, 820) **caracterizado por**:

um mecanismo de reconhecimento de voz (824) que tem acesso a:

um corpus de treinamento de expressões de voz digitalizadas de classes conhecidas;

uma pluralidade de classificadores de átona, em que cada classificador de átona compreende uma função de decisão para determinar a presença de um evento dentro do corpus de treinamento; e

um detector de conjunto compreendendo uma pluralidade dos classificadores de átona que juntos são melhores para determinar a presença de um evento de sinal de voz do que qualquer um dos classificadores de átona constituintes;

em que o mecanismo de reconhecimento de voz (824) compreende um extrator de evento (804, 825, 854) para extrair eventos de sinal de voz e padrões dos eventos de sinal de voz do sinal de voz digital (801, 821, 851), em que os eventos de sinal de voz e os padrões dos eventos de sinal de voz são relevantes no reconhecimento de voz;

em que o mecanismo de reconhecimento de voz (824) compreende pelo menos um processador que está configurado para realizar uma pluralidade de operações, em que a pluralidade de operações compreende:

detectar localizações de eventos de sinal de voz relevantes no sinal de voz digital (801, 821, 851), em que cada um dos eventos de sinal de voz compreende informação espectral e informação temporal;

captura de características espectrais e relações temporais entre todos os eventos de sinal de voz;

segmentar o sinal de voz digital (801, 821, 851) com base nas localizações detectadas dos eventos de sinal de voz detectados;

analisar o sinal de voz digital segmentado (801, 821, 851), em que a análise é sincronizada com os eventos do sinal de voz;

detectar padrões no sinal de voz digital (801, 821, 851) com a informação espectral capturada, as relações temporais e o sinal de voz digital analisado (801, 821, 851);

fornecer uma lista de alternativas perceptivas para dados de voz reconhecidos que correspondem aos padrões detectados no sinal de voz digital (801, 821, 851); e

eliminar a ambiguidade entre as alternativas perceptivas para os dados de voz reconhecidos com base na análise de um ou mais dos eventos de sinal de voz para melhorar os dados de voz reconhecidos;

em que pelo menos um processador é configurado para realizar uma ou mais das operações usando o detector de conjunto; e

um módulo acoplado ao mecanismo de reconhecimento de voz (824), em que o módulo é configurado para emitir os dados de voz reconhecidos aprimorados (811, 841).

2. Sistema (800, 820), de acordo com a reivindicação 1, **caracterizado pelo fato de** que ainda compreende um mecanismo para iniciar pelo menos uma ação em resposta a pelo menos uma parte dos dados de voz reconhecidos aprimorados de saída (811, 841).

3. Sistema (800, 820), de acordo com a reivindicação 2, **caracterizado pelo fato de** que pelo menos uma ação compreende qualquer um dentre:

uma conversão dos dados de voz reconhecidos aprimorados (811, 841) em pelo menos uma sequência de texto; ou

uma supressão de uma saída de áudio quando certas palavras são detectadas.

4. Sistema (800, 820), de acordo com a reivindicação 2, **caracterizado pelo fato de** que ainda compreende um mecanismo para detectar pelo menos um comando nos dados de voz reconhecidos aprimorados (811, 841);

em que a pelo menos uma ação compreende um início de resposta ao comando detectado.

5. Sistema (800, 820), de acordo com a reivindicação 1, **caracterizado pelo fato de** que compreende adicionalmente:

o corpus de treinamento de expressões de voz digitalizada de classes conhecidas;

em que o pelo menos um processador é ainda configurado para:

estabilizar a pluralidade de classificadores de átono; e

construir o detector de conjunto.

6. Sistema (800, 820), de acordo com a reivindicação 5, **caracterizado pelo fato de** que o pelo menos um processador é configurado para construir iterativamente o detector de conjunto com um algoritmo de reforço para formar um detector de conjunto reforçado.

7. Sistema (800, 820), de acordo com a reivindicação 6, **caracterizado pelo fato de** que o pelo menos um processador é configurado para simplificar o detector de conjunto reforçado construído.

8. Sistema (800, 820), de acordo com a reivindicação 7, **caracterizado pelo fato de** que o pelo menos um processador é configurado para converter o detector de conjunto reforçado construído simplificado em um detector em cascata.

9. Sistema (800, 820), de acordo com a reivindicação 1, **caracterizado pelo fato de** que a lista das alternativas perceptivas para os dados de voz reconhecidos compreende uma pluralidade de grupos perceptivos.

10. Sistema (800, 820), de acordo com a reivindicação 1, **caracterizado pelo fato de** que pelo menos um processador é adicionalmente configurado para rejeitar uma ou mais regiões do sinal de voz digital (801, 821, 851) que não contêm um ou mais dos eventos de sinal de voz .

11. Sistema (800, 820), de acordo com a reivindicação 1, **caracterizado pelo fato de** que o pelo menos um processador é adicionalmente configurado para detectar sequências dos eventos de sinal de voz com base nos padrões detectados.

12. Sistema (800, 820), de acordo com a reivindicação 1, **caracterizado pelo fato de** que o pelo menos um processador é ainda configurado para reconhecer pistas alternativas de voz para fortalecer o reconhecimento.

13. Sistema (800, 820), de acordo com a reivindicação 1, **caracterizado pelo fato de** que compreende adicionalmente:

um filtro de pré-segmentação; e

um extrator de característica;

em que o filtro de pré-segmentação é configurado para definir intervalos que são usados para sincronizar computações de características;

em que a segmentação do sinal de voz digital (801, 821, 851) é baseada em diferenças perceptuais dos intervalos definidos; e

em que o extrator de característica é configurado para extrair características relativas aos eventos de sinal de voz a partir do sinal de voz digital (801, 821, 851) segmentado.

14. Sistema (800, 820), de acordo com a reivindicação 1, **caracterizado pelo fato de** que o pelo menos um processador é ainda configurado para

converter os dados de voz reconhecidos aprimorados (811, 841) em pelo menos uma sequência de texto; e

inserir pontuação automaticamente em pelo menos uma sequência de texto.

15. Método de reconhecimento de voz **caracterizado pelo fato de** que compreende:

acessar uma pluralidade de classificadores de átona, em que cada classificador de átona compreende uma função de decisão para determinar a presença de um evento dentro de um corpus de treinamento de expressões de voz digitalizadas de classes conhecidas; e

um detector de conjunto compreendendo uma pluralidade dos classificadores de átona, que juntos são melhores na determinação da presença de um evento de sinal de voz do que qualquer um dos classificadores de átona constituintes;

receber um sinal de voz;

digitalizar o sinal de voz recebido;

detectar localizações de eventos de sinal de voz relevantes no sinal de voz recebido e digitalizado, em que cada um dos eventos de sinal de voz relevantes compreende informação espectral e informação temporal;

capturar características espectrais de e relações temporais entre todos os eventos de sinal de voz;

segmentar o sinal de voz recebido e digitalizado com base nas localizações detectadas dos eventos de sinal de voz:

analisar o sinal de voz segmentado, recebido e digitalizado, em que a análise é sincronizada com os eventos do sinal de voz;

detectar padrões no sinal de voz digitalizado com a informação espectral capturada, as relações temporais e o sinal de voz analisado;

reconhecer dados de voz que correspondem ao sinal de voz digitalizado analisado, em que a etapa de reconhecimento dos da-

dos de voz compreende as etapas de:

fornecer uma lista de alternativas perceptivas para os dados de voz reconhecidos que correspondem aos padrões detectados no sinal de voz digitalizado; e

eliminar a ambiguidade entre as alternativas perceptivas para os dados de voz reconhecidos com base na análise de um ou mais dos eventos de sinal de voz para melhorar os dados de voz reconhecidos; e a saída dos dados de voz aprimorados.

16. Método, de acordo com a reivindicação 15, **caracterizado pelo fato de** que adicionalmente compreende as etapas de:

estabelecer a pluralidade de classificadores de átona; e

construir o detector de conjunto;

em que a etapa de construção de um detector de conjunto compreende as etapas de

armazenar uma pluralidade de sinais de voz, em que os sinais de voz compreendem exemplos de treinamento armazenados em um sistema de reconhecimento automático de voz,

extrair padrões de eventos de uma pluralidade de exemplos de treinamento armazenados, em que os padrões de evento compreendem localizações de características distintas na pluralidade de sinais de voz armazenados, e

executar iterativamente as etapas de

acessar uma amostra da pluralidade de sinais de voz tendo padrões de eventos correspondentes,

alinhar eventos de sinais de voz individuais entre as amostras, em que alinhar compreende o alinhamento dos eventos dos sinais de voz individuais temporalmente com base nos padrões de eventos correspondentes,

avaliar a eficácia de uma pluralidade de detectores de átona na detecção dos padrões de evento,

aplicar um esquema de ponderação à pluralidade de detectores de átona com base na eficácia relativa dos detectores de átona, em que os detectores de átona mais eficazes têm os pesos mais elevados; e

adicionar pelo menos um detector de átona adicional à pluralidade de detectores de átona:

em que a iteração é realizada até que a eficácia do esquema de ponderação atinja um determinado padrão de eficiência para detectar os padrões de evento.

17. Método, de acordo com a reivindicação 16, **caracterizado pelo fato de** que a etapa de acessar uma amostra da pluralidade de sinais de voz que tem padrões de eventos correspondentes compreende adicionalmente a etapa de identificar regiões automaticamente na pluralidade de sinais de voz que contenham os padrões de eventos, que compreende as etapas de:

alinhar a pluralidade de sinais de voz relativos a um tempo comum;

projetar uma ou mais localizações de eventos dos sinais de voz individuais no eixo de tempo comum; e

identificar regiões no eixo do tempo tendo uma concentração dos locais de evento na forma de regiões na pluralidade de sinais de voz que contêm os padrões de eventos.

18. Método, de acordo com a reivindicação 15, **caracterizado pelo fato de** que a etapa de acessar uma amostra da pluralidade de sinais de voz que tem padrões de eventos correspondentes compreende adicionalmente a etapa de identificar regiões automaticamente na pluralidade de sinais de voz que contenham os ditos padrões de eventos, que compreende as etapas de:

acessar um conjunto de treinamento;

converter o sinal de voz dentro de regiões do espaço tem-

po-trajetória que contém todos os eventos de sinal de voz dos exemplos de treinamento positivo; e

realizar repetidamente as etapas de:

calcular as contagens de exemplos negativos para todas as regiões do espaço tempo-trajetória;

selecionar uma região das regiões do espaço tempo-trajetória com o menor número de eventos de exemplos de treinamento negativo; e

eliminar exemplos negativos sem eventos de sinal de voz na região selecionada de consideração posterior;

até que seja criada uma cascata que funcione perfeitamente no referido conjunto de treinamento.

19. Sistema para reconhecimento de voz (800, 820) que corresponde a um sinal de voz digital (801, 821, 851), o sistema (800, 820) **caracterizado por**:

um mecanismo de reconhecimento de voz (824) que tem acesso a:

um corpus de treinamento de expressões de voz digitalizadas de classes conhecidas;

uma pluralidade de classificadores de átona, em que cada classificador de átona compreende uma função de decisão para determinar a presença de um evento dentro do corpus de treinamento; e

um detector de conjunto compreendendo uma pluralidade dos classificadores de átona, que juntos são melhores na determinação da presença de um evento de sinal de voz do que qualquer um dos classificadores de átona constituintes;

em que o mecanismo de reconhecimento de voz (824) compreende um extrator de eventos (804, 825, 854) para extrair eventos de sinal de voz e padrões dos eventos de sinal de voz do sinal de voz digital (801, 821, 851), em que os eventos de sinal de voz e pa-

drões dos eventos de sinal de voz são relevantes no reconhecimento de voz, em que o mecanismo de reconhecimento de voz (824) compreende pelo menos um processador que está configurado para executar uma pluralidade de operações, em que a pluralidade de operações compreende:

detectar localizações de eventos de sinal de voz relevantes no sinal de voz digital (801, 821, 851), em que cada um dos eventos de sinal de voz compreende informação espectral e informação temporal;

capturar características espectrais e relações temporais entre todos os eventos de sinal de voz;

segmentar o sinal de voz digital (801, 821, 851) com base nas localizações detectadas dos eventos de sinal de voz detectados;

analisar o sinal de voz digital segmentado, em que a análise é sincronizada com os eventos do sinal de voz;

detectar padrões no sinal de voz digital (801, 821, 851) com a informação espectral capturada, as relações temporais e o sinal de voz digital analisado;

fornecer uma lista de alternativas perceptivas para dados de voz reconhecidos que correspondem aos padrões detectados no sinal de voz digital; e

eliminar a ambiguidade entre as alternativas perceptivas para os dados de voz reconhecidos com base na análise de um ou mais dos eventos de sinal de voz para melhorar os dados de voz reconhecidos;

em que o pelo menos um processador é configurado para realizar uma ou mais das operações usando o detector de conjunto;

um módulo acoplado ao mecanismo de reconhecimento de voz (824), em que o módulo é configurado para emitir os dados de voz reconhecidos aprimorados (811, 841);

um mecanismo de pontuação automática acoplado com um banco de dados que contém dados de treinamento, em que o mecanismo de pontuação automática compreende pelo menos um processador estatístico para adicionar a pontuação aos dados de voz reconhecidos aprimorados (811, 841) com o uso dos dados de treinamento na forma de texto pontuados baseados em estatística;

um pontuador baseado em regras acoplado com um banco de dados de regras lexicais, em que o dito pontuador baseado em regras adiciona pontuação aos dados de voz reconhecidos aprimorados (811, 841) usando regras do banco de dados de regras lexicais na forma de texto pontuado baseado em regras; e

um módulo de decisão para determinar se o texto pontuado ou texto pontuado baseado em estatística produz um resultado pontuado melhor; e

um mecanismo que é configurado para produzir o melhor resultado pontuado, com base na determinação.

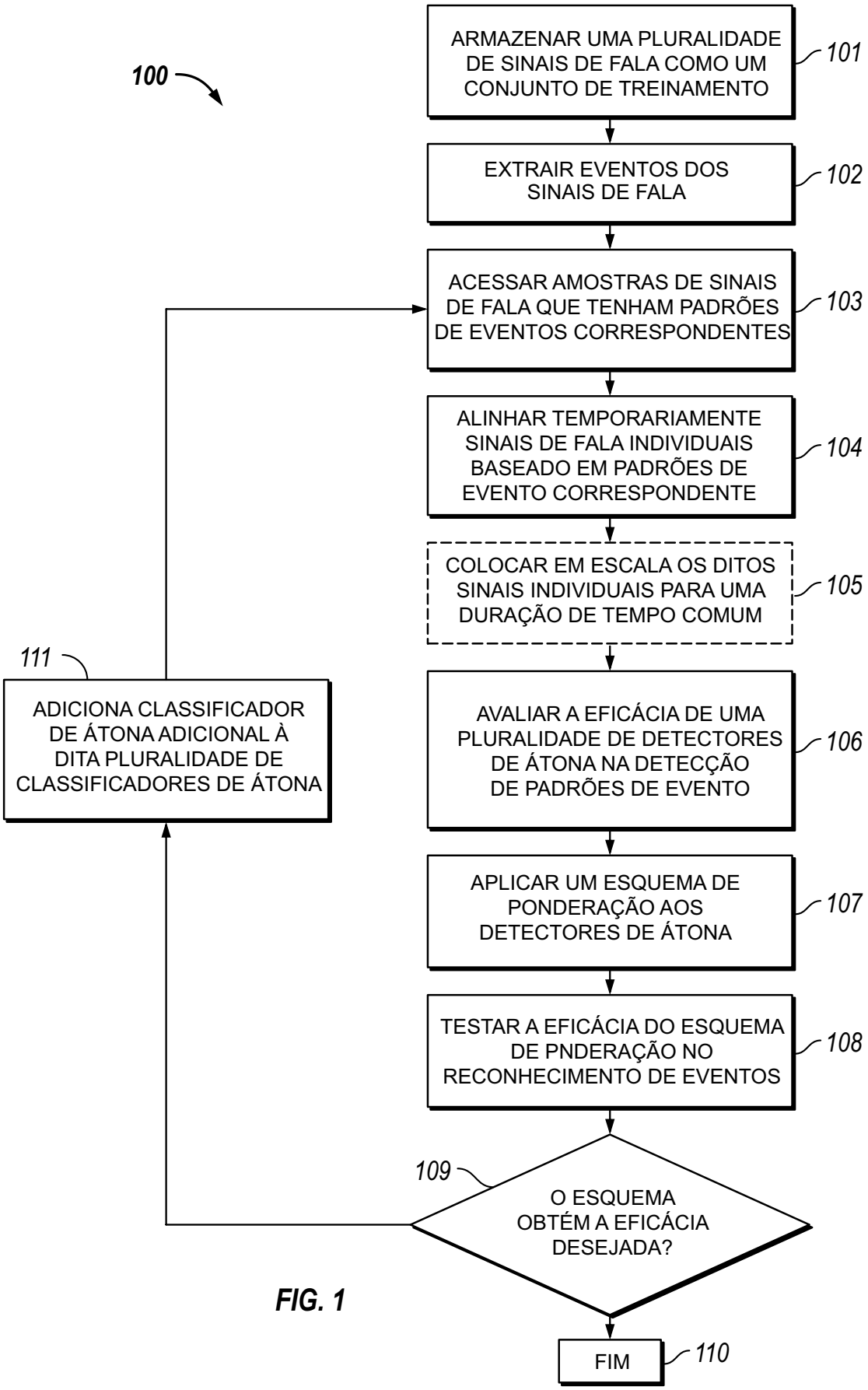
100
ARMAZENAR UMA PLURALIDADE DE SINAIS DE FALA COMO UM CONJUNTO DE TREINAMENTO
101
EXTRAIR EVENTOS DOS SINAIS DE FALA
102
ACESSAR AMOSTRAS DE SINAIS DE FALA QUE TENHAM PADRÕES DE EVENTOS CORRESPONDENTES
103
ALINHAR TEMPORARIAMENTE SINAIS DE FALA INDIVIDUAIS BASEADO EM PADRÕES DE EVENTO CORRESPONDENTE
104
COLOCAR EM ESCALA OS DITOS SINAIS INDIVIDUAIS PARA UMA DURAÇÃO DE TEMPO COMUM
105
111
ADICIONA CLASSIFICADOR DE ÁTONA ADICIONAL À DITA PLURALIDADE DE CLASSIFICADORES DE ÁTONA
AVALIAR A EFICÁCIA DE UMA PLURALIDADE DE DETECTORES DE ÁTONA NA DETECÇÃO DE PADRÕES DE EVENTO
106
APLICAR UM ESQUEMA DE PONDERAÇÃO AOS DETECTORES DE ÁTONA
107
TESTAR A EFICÁCIA DO ESQUEMA DE PNDERAÇÃO NO RECONHECIMENTO DE EVENTOS
108
109
O ESQUEMA OBTÉM A EFICÁCIA DESEJADA?
FIM
110

*FIG. 1*

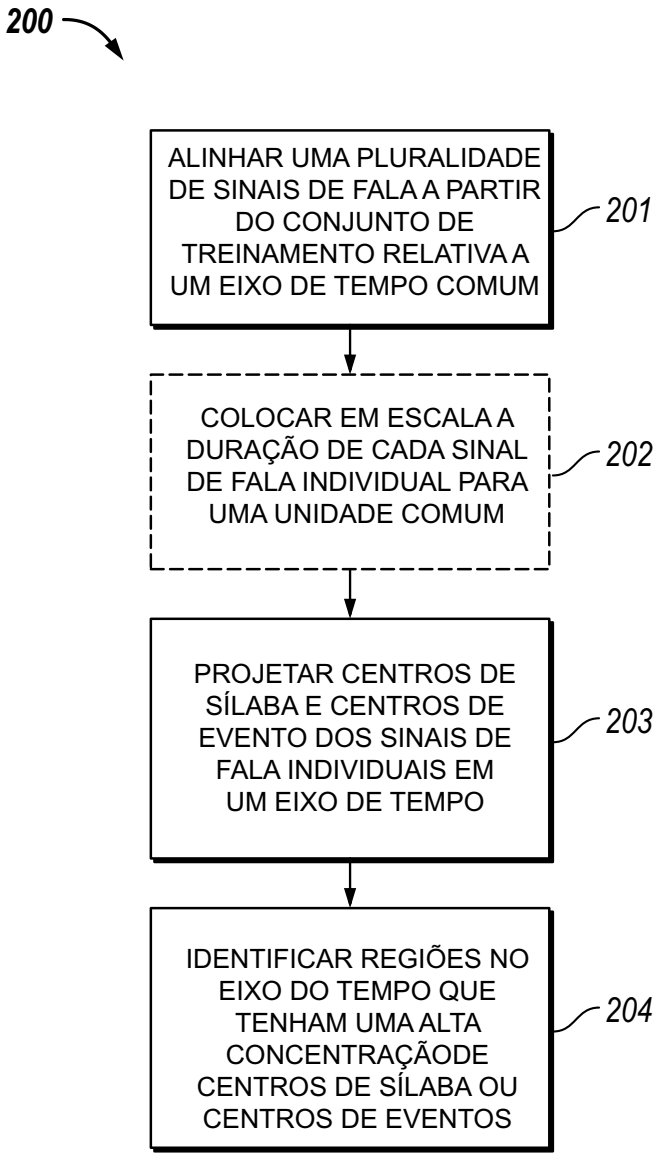
200
ALINHAR UMA PLURALIDADE DE SINAIS DE FALA A PARTIR DO CONJUNTO DE TREINAMENTO RELATIVA A UM EIXO DE TEMPO COMUM
201
COLOCAR EM ESCALA A DURAÇÃO DE CADA SINAL DE FALA INDIVIDUAL PARA UMA UNIDADE COMUM
202
PROJETAR CENTROS DE SÍLABA E CENTROS DE EVENTO DOS SINAIS DE FALA INDIVIDUAIS EM UM EIXO DE TEMPO
203
IDENTIFICAR REGIÕES NO EIXO DO TEMPO QUE TENHAM UMA ALTA CONCENTRAÇÃODE CENTROS DE SÍLABA OU CENTROS DE EVENTOS
204

*FIG. 2*

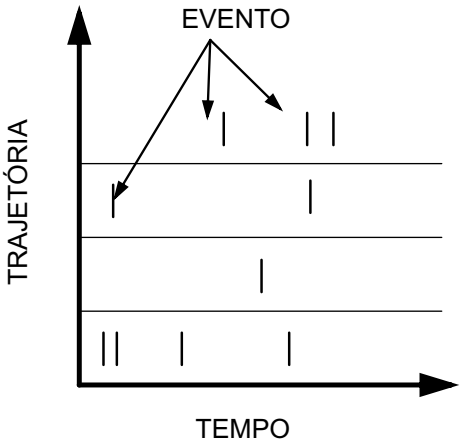


FIG. 3A

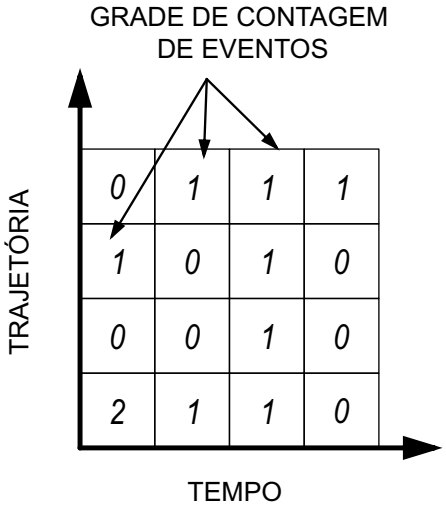


FIG. 3B

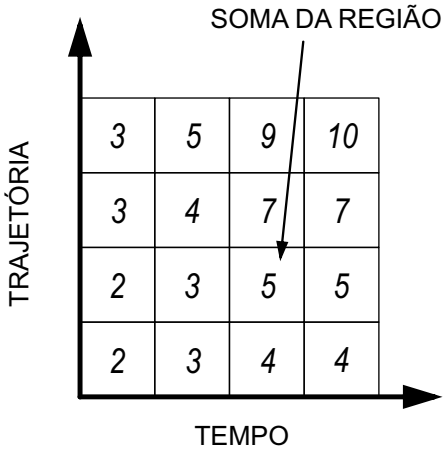


FIG. 3C

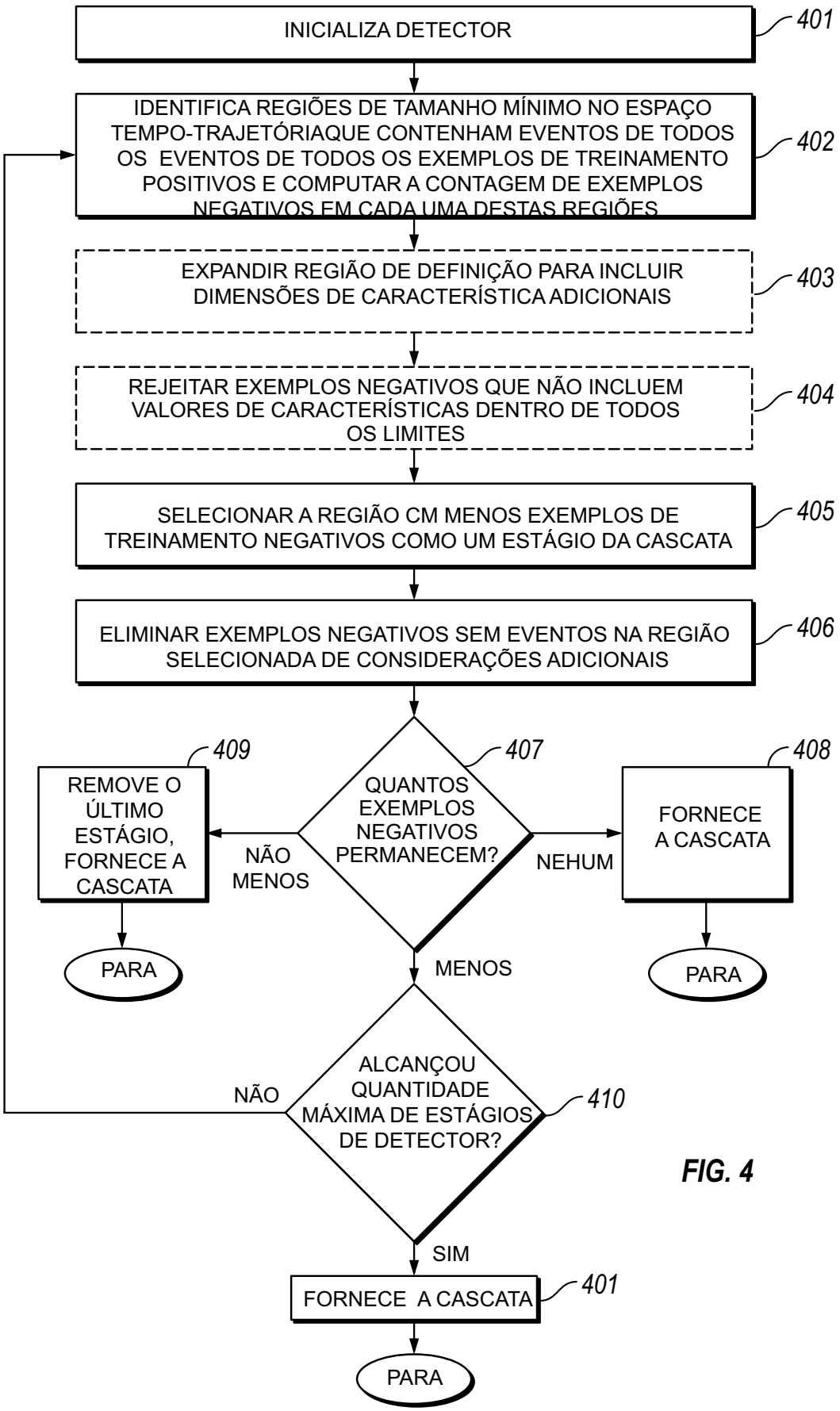
INICIALIZA DETECTOR
401
IDENTIFICA REGIÕES DE TAMANHO MÍNIMO NO ESPAÇO TEMPO-TRAJETÓRIAQUE CONTENHAM EVENTOS DE TODOS OS EVENTOS DE TODOS OS EXEMPLOS DE TREINAMENTO POSITIVOS E COMPUTAR A CONTAGEM DE EXEMPLOS NEGATIVOS EM CADA UMA DESTAS REGIÕES
402
EXPANDIR REGIÃO DE DEFINIÇÃO PARA INCLUIR DIMENSÕES DE CARACTERÍSTICA ADICIONAIS
403
REJEITAR EXEMPLOS NEGATIVOS QUE NÃO INCLUEM VALORES DE CARACTERÍSTICAS DENTRO DE TODOS OS LIMITES
404
SELECIONAR A REGIÃO CM MENOS EXEMPLOS DE TREINAMENTO NEGATIVOS COMO UM ESTÁGIO DA CASCATA
405
ELIMINAR EXEMPLOS NEGATIVOS SEM EVENTOS NA REGIÃO SELECIONADA DE CONSIDERAÇÕES ADICIONAIS
406
QUANTOS EXEMPLOS NEGATIVOS PERMANECEM?
407
NÃO MENOS
NEHUM
REMOVE O ÚLTIMO ESTÁGIO, FORNECE A CASCATA
409
FORNECE A CASCATA
408
PARA
PARA
MENOS
ALCANÇOU QUANTIDADE MÁXIMA DE ESTÁGIOS DE DETECTOR?
410
NÃO
SIM
FORNECE A CASCATA
401
PARA

*FIG. 4*

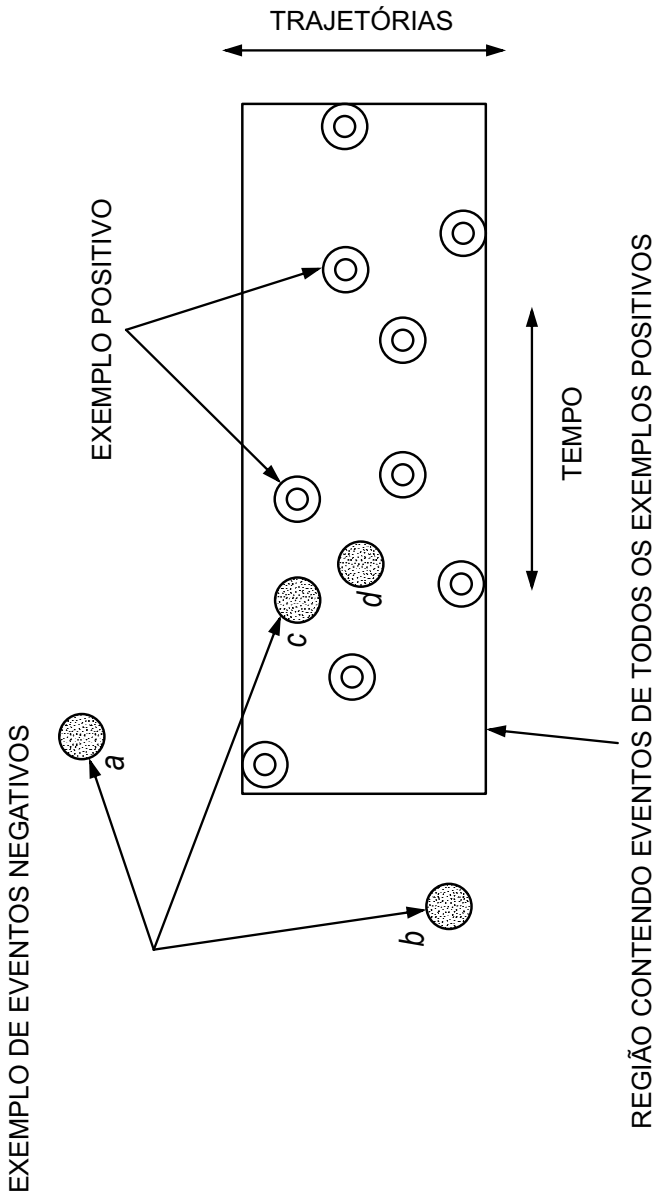
TRAJETÓRIAS
EXEMPLO POSITIVO
TEMPO
EXEMPLO DE EVENTOS NEGATIVOS
REGIÃO CONTENDO EVENTOS DE TODOS OS EXEMPLOS POSITIVOS
a
b
c
d

*FIG. 5*

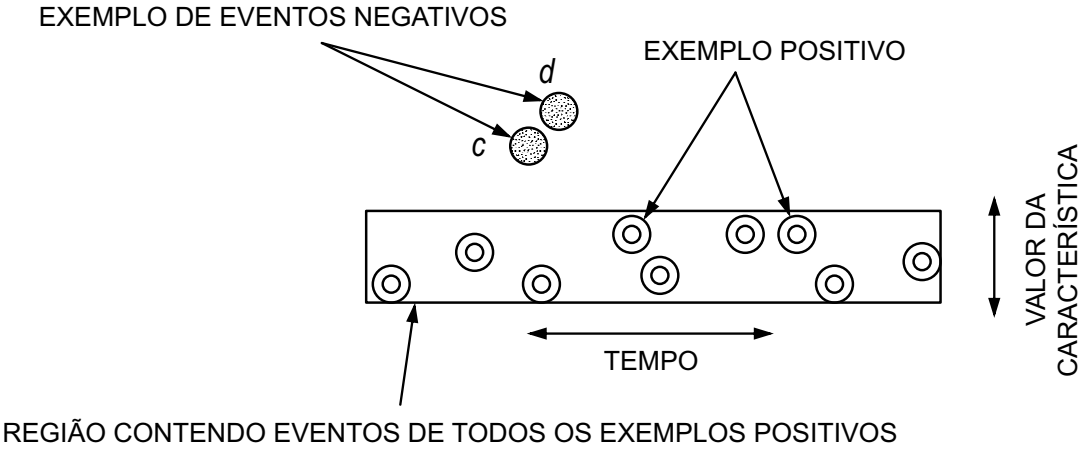
EXEMPLO DE EVENTOS NEGATIVOS
EXEMPLO POSITIVO
d
c
VALOR DA CARACTERÍSTICA
TEMPO
REGIÃO CONTENDO EVENTOS DE TODOS OS EXEMPLOS POSITIVOS

***FIG. 6A***

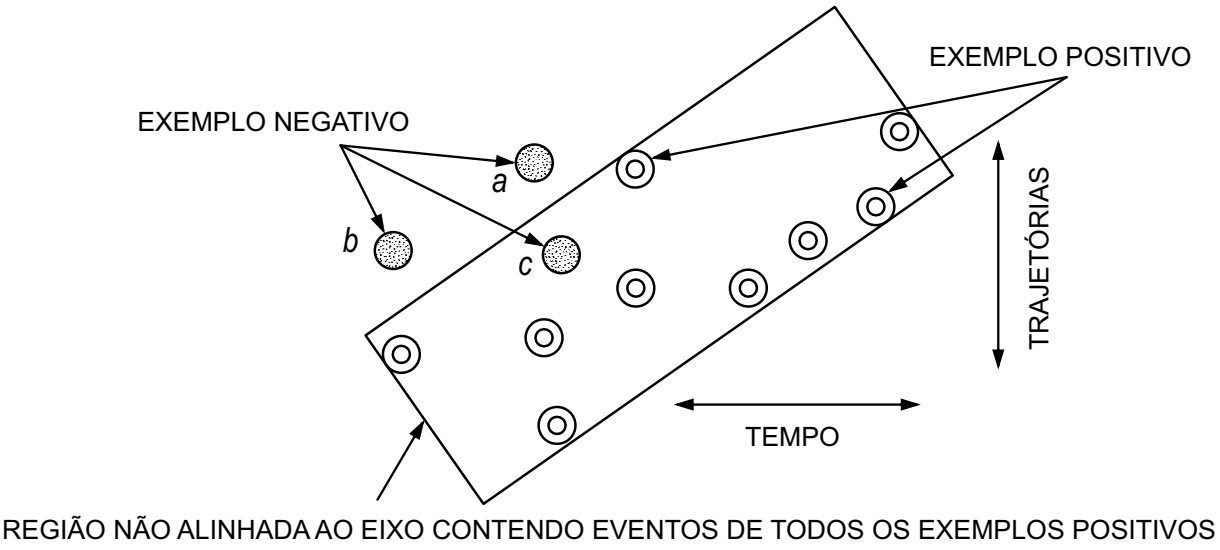
EXEMPLO POSITIVO
EXEMPLO NEGATIVO
a
b
c
TRAJETÓRIAS
TEMPO
REGIÃO NÃO ALINHADA AO EIXO CONTENDO EVENTOS DE TODOS OS EXEMPLOS POSITIVOS

***FIG. 6B***

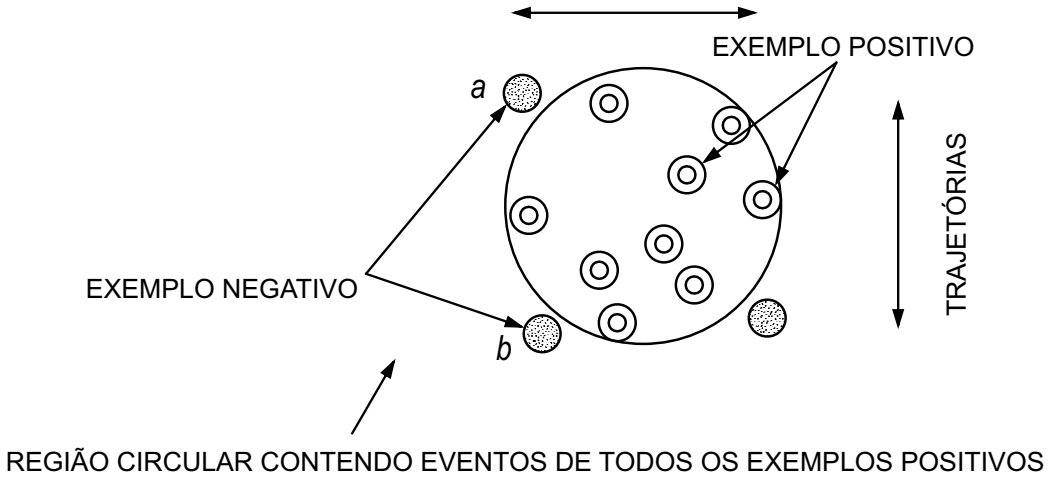
EXEMPLO POSITIVO
a
TRAJETÓRIAS
EXEMPLO NEGATIVO
b
REGIÃO CIRCULAR CONTENDO EVENTOS DE TODOS OS EXEMPLOS POSITIVOS

*FIG. 6C*

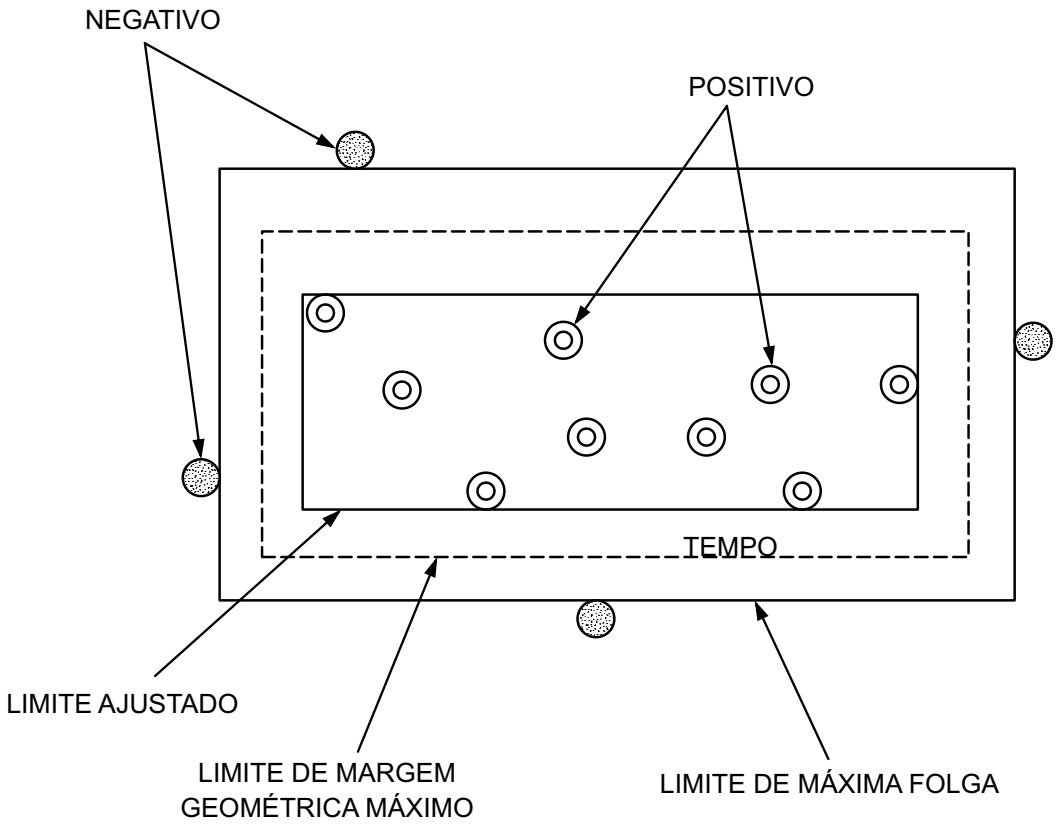
NEGATIVO
POSITIVO
TEMPO
LIMITE AJUSTADO
LIMITE DE MARGEM
GEOMÉTRICA MÁXIMO
LIMITE DE MÁXIMA FOLGA

*FIG. 7*

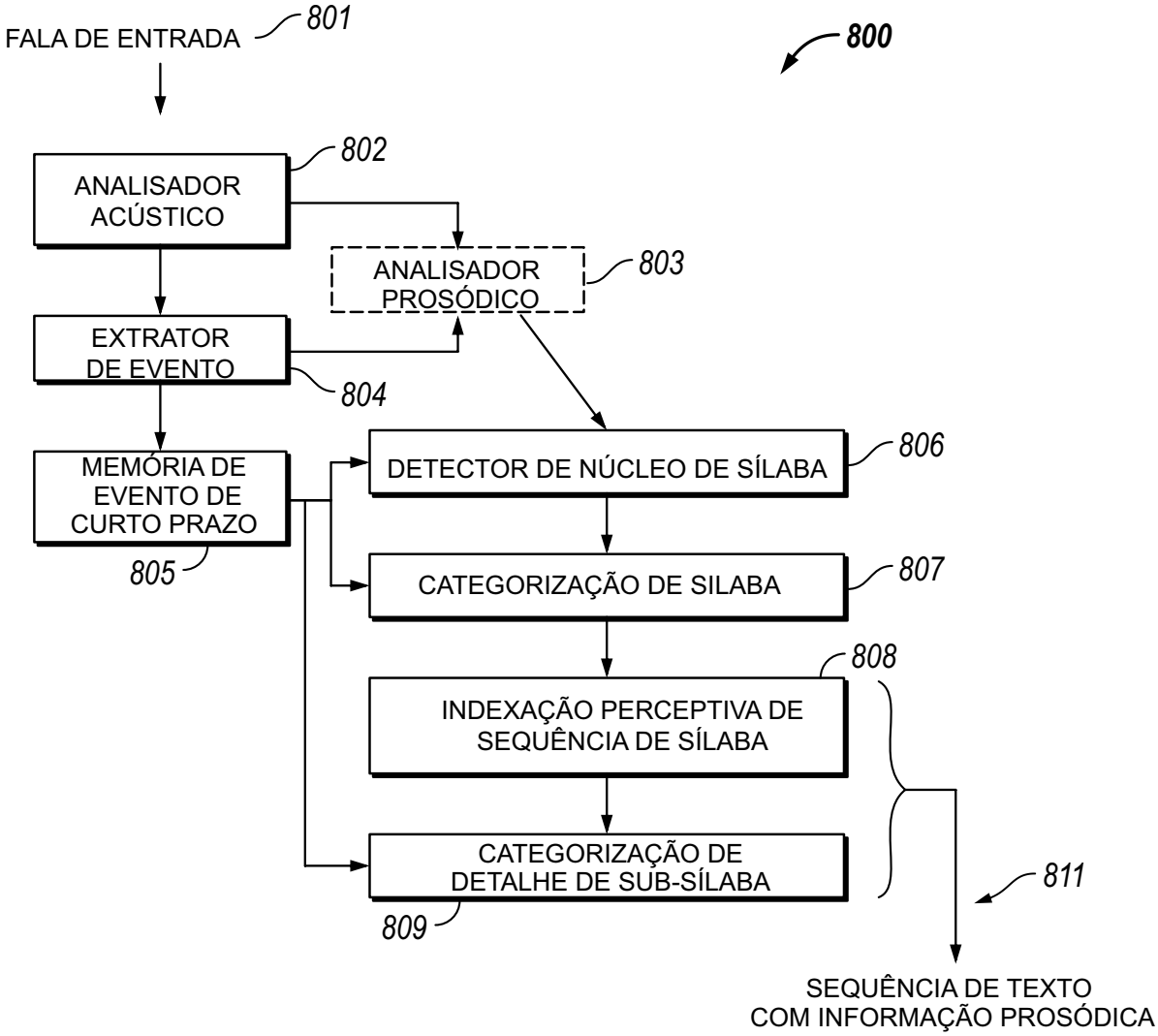
FALA DE ENTRADA
801
800
ANALISADOR ACÚSTICO
802
ANALISADOR PROSÓDICO
803
EXTRATOR DE EVENTO
804
MEMÓRIA DE EVENTO DE CURTO PRAZO
805
DETECTOR DE NÚCLEO DE SÍLABA
806
CATEGORIZAÇÃO DE SILABA
807
INDEXAÇÃO PERCEPTIVA DE SEQUÊNCIA DE SÍLABA
808
CATEGORIZAÇÃO DE DETALHE DE SUB-SÍLABA
809
811
SEQUÊNCIA DE TEXTO COM INFORMAÇÃO PROSÓDICA

*FIG. 8A*

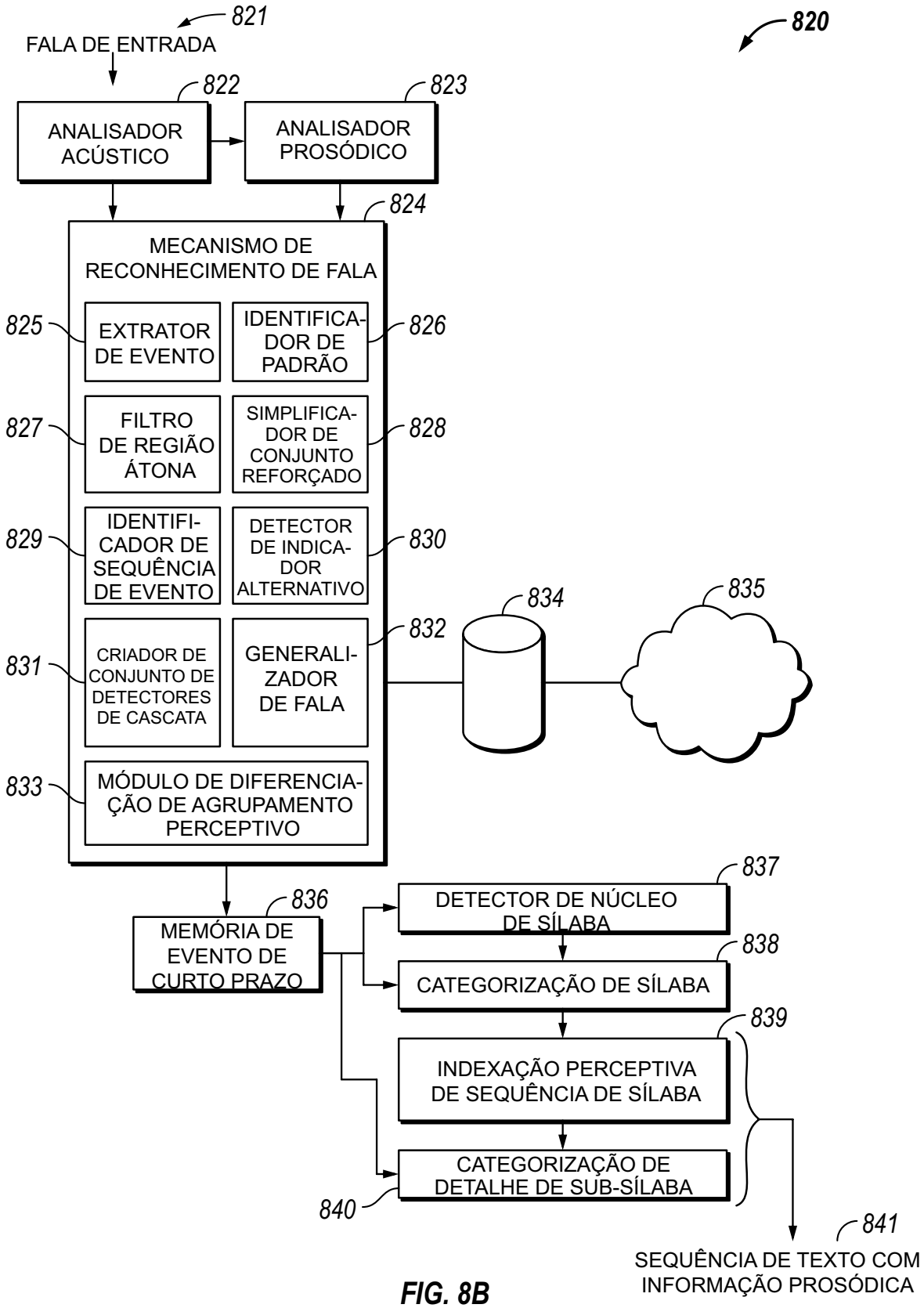
821
FALA DE ENTRADA
820
822
ANALISADOR ACÚSTICO
823
ANALISADOR PROSÓDICO
824
MECANISMO DE RECONHECIMENTO DE FALA
825
EXTRATOR DE EVENTO
826
IDENTIFICA-DOR DE PADRÃO
827
FILTRO DE REGIÃO ÁTONA
828
SIMPLIFICA-DOR DE CONJUNTO REFORÇADO
829
IDENTIFI-CADOR DE SEQUÊNCIA DE EVENTO
830
DETECTOR DE INDICA-DOR ALTERNATIVO
831
CRIADOR DE CONJUNTO DE DETECTORES DE CASCATA
832
GENERALI-ZADOR DE FALA
833
MÓDULO DE DIFERENCIA-ÇÃO DE AGRUPAMENTO PERCEPTIVO
834
835
836
MEMÓRIA DE EVENTO DE CURTO PRAZO
837
DETECTOR DE NÚCLEO DE SÍLABA
838
CATEGORIZAÇÃO DE SÍLABA
839
INDEXAÇÃO PERCEPTIVA DE SEQUÊNCIA DE SÍLABA
840
CATEGORIZAÇÃO DE DETALHE DE SUB-SÍLABA
841
SEQUÊNCIA DE TEXTO COM INFORMAÇÃO PROSÓDICA

*FIG. 8B*

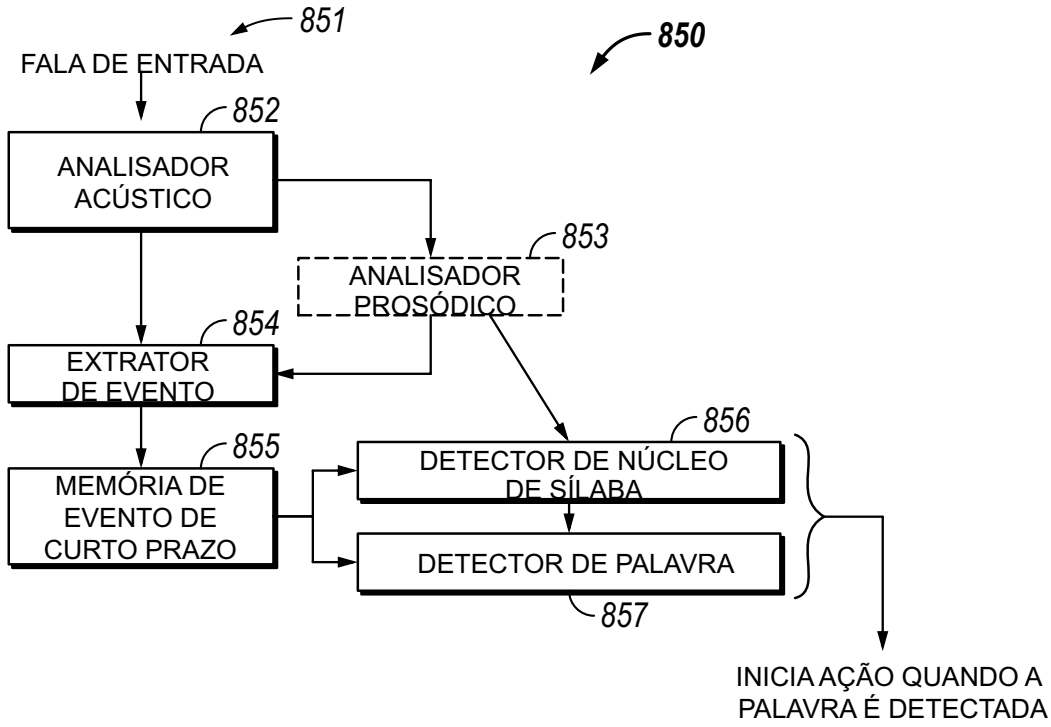
851
850
FALA DE ENTRADA
852
ANALISADOR ACÚSTICO
853
ANALISADOR PROSÓDICO
854
EXTRATOR DE EVENTO
855
MEMÓRIA DE EVENTO DE CURTO PRAZO
856
DETECTOR DE NÚCLEO DE SÍLABA
DETECTOR DE PALAVRA
857
INICIA AÇÃO QUANDO A PALAVRA É DETECTADA

*FIG. 8C*

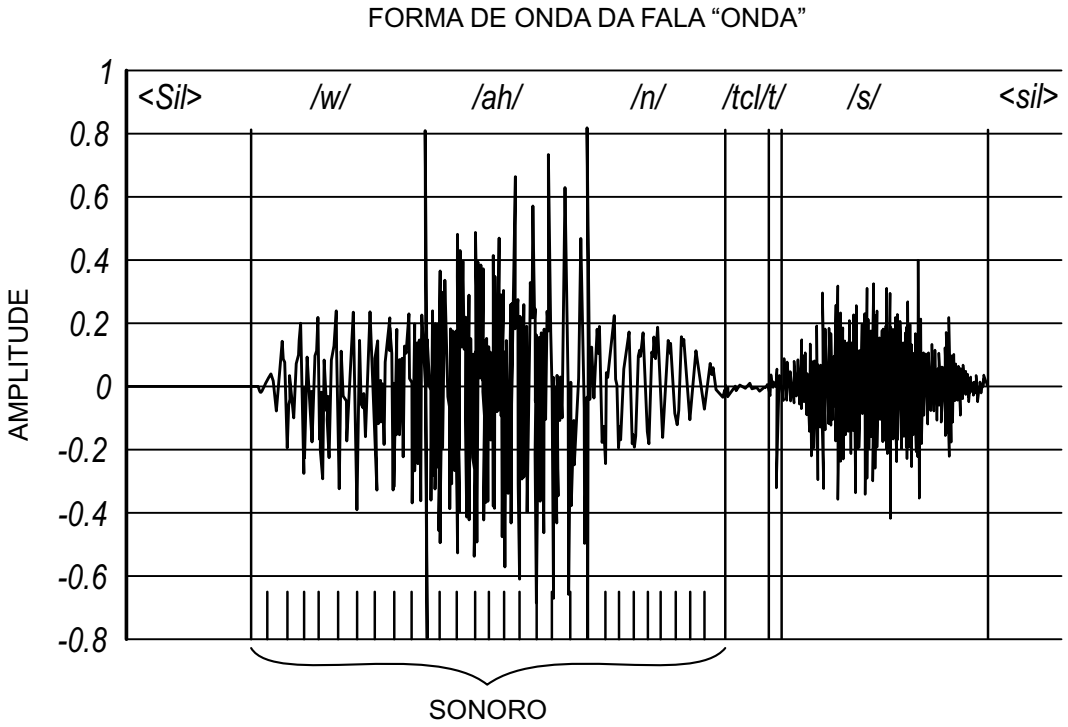
FORMA DE ONDA DA FALA “ONDA”
AMPLITUDE
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
<Sil>
/w/
/ah/
/n/
/tcl/t/
/s/
<sil>
SONORO

*FIG. 9*

FORMULA DE CONTRASTE PERCEPTIVO

$$C_{AB} = \frac{(A_{\text{MÉDIO}} - B_{\text{MÉDIO}})}{(A_{\text{MÉDIO}} + B_{\text{MÉDIO}} + \varepsilon)};$$

ONDE $A_{\text{MÉDIO}}$ E $B_{\text{MÉDIO}}$ SÃO VALORE DE INTERVALO MÉDIO. PARÂMETRO ε É O VALOR DE NÍVEL DE ATIVAÇÃO PERCEPTIVA MÍNIMA

***FIG. 10***

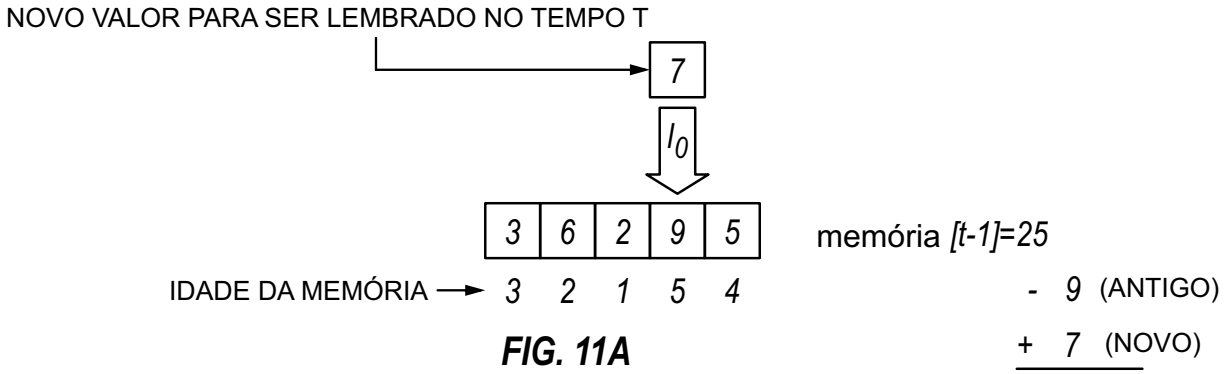


***FIG. 11A***

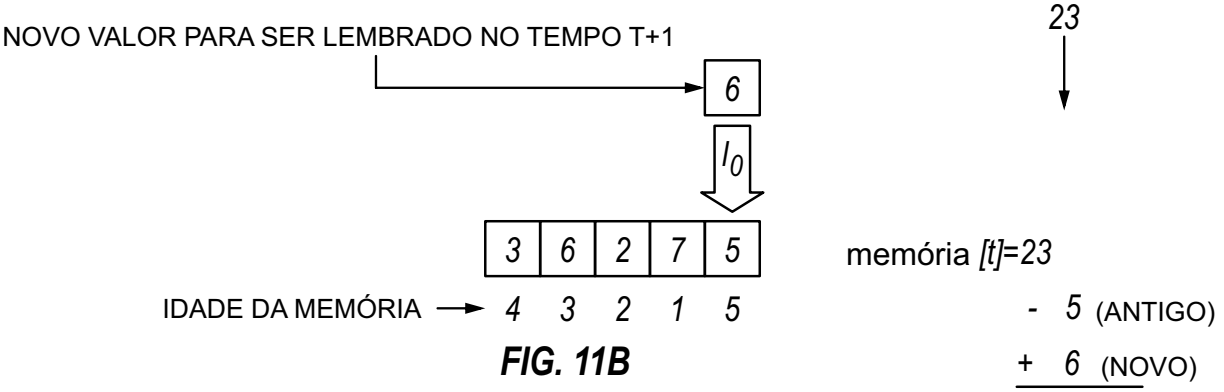


***FIG. 11B***

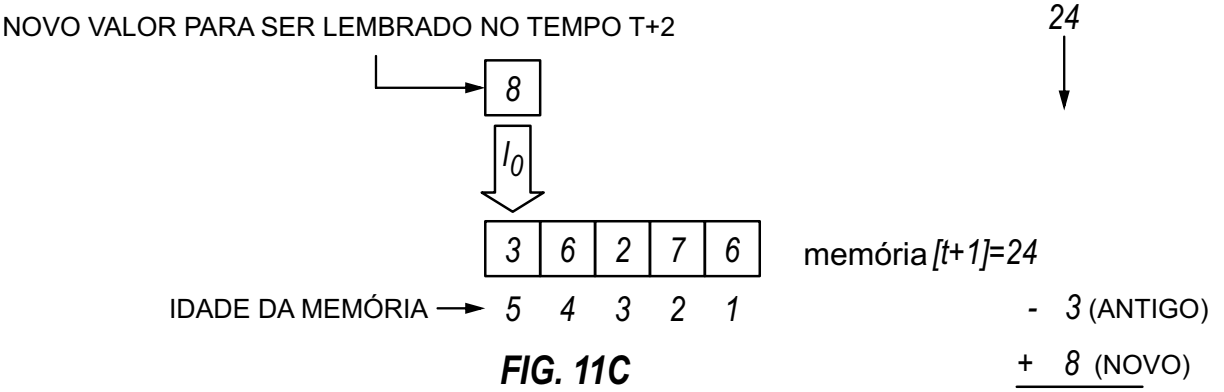


***FIG. 11C***

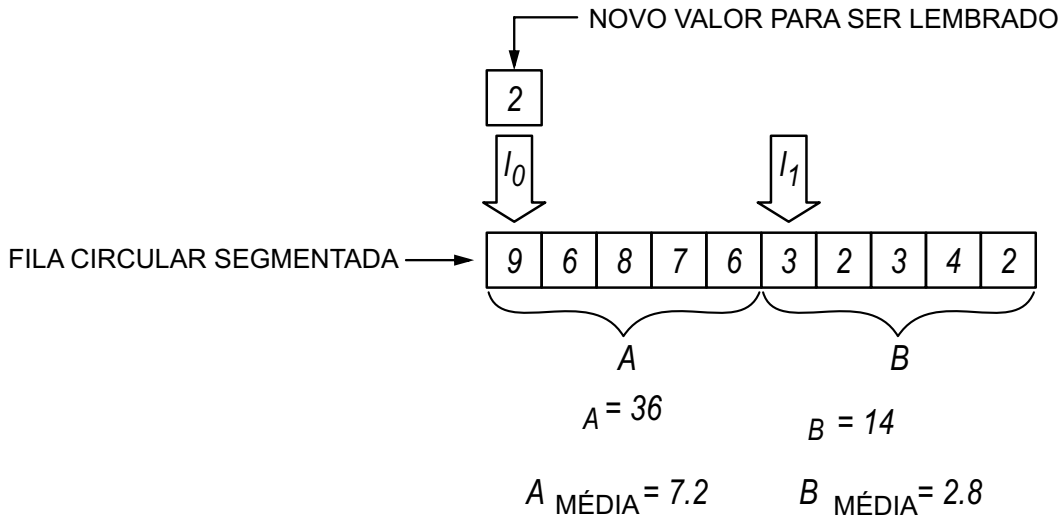
NOVO VALOR PARA SER LEMBRADO
2
$I_0$
$I_1$
FILA CIRCULAR SEGMENTADA
9 6 8 7 6 3 2 3 4 2
A
B
$A = 36$
$B = 14$
$A_{MÉDIA} = 7.2$
$B_{MÉDIA} = 2.8$

***FIG. 12***

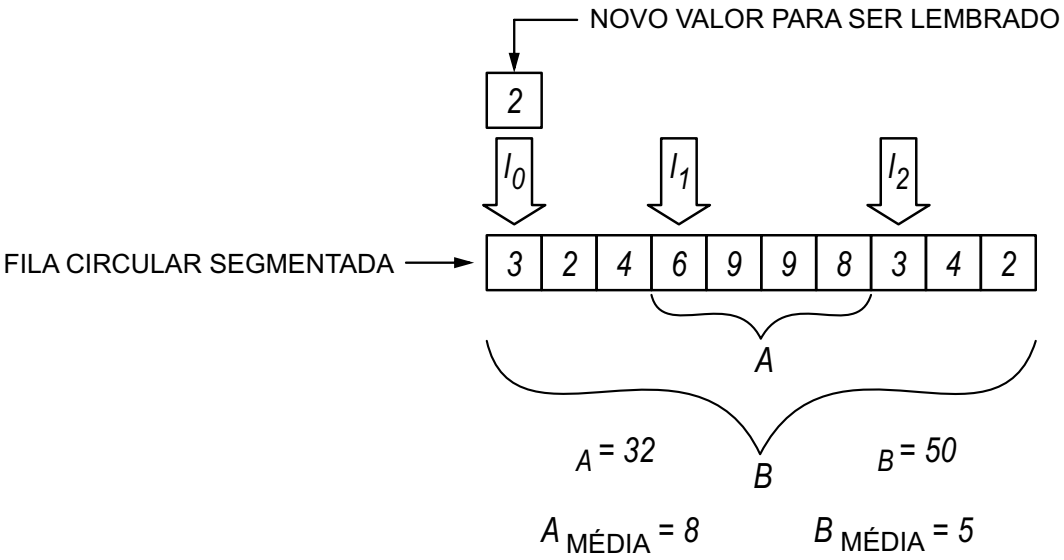
NOVO VALOR PARA SER LEMBRADO
2
$I_0$
$I_1$
$I_2$
FILA CIRCULAR SEGMENTADA
3 2 4 6 9 9 8 3 4 2
A
B
$A = 32$
$B = 50$
$A_{MÉDIA} = 8$
$B_{MÉDIA} = 5$

***FIG. 13***

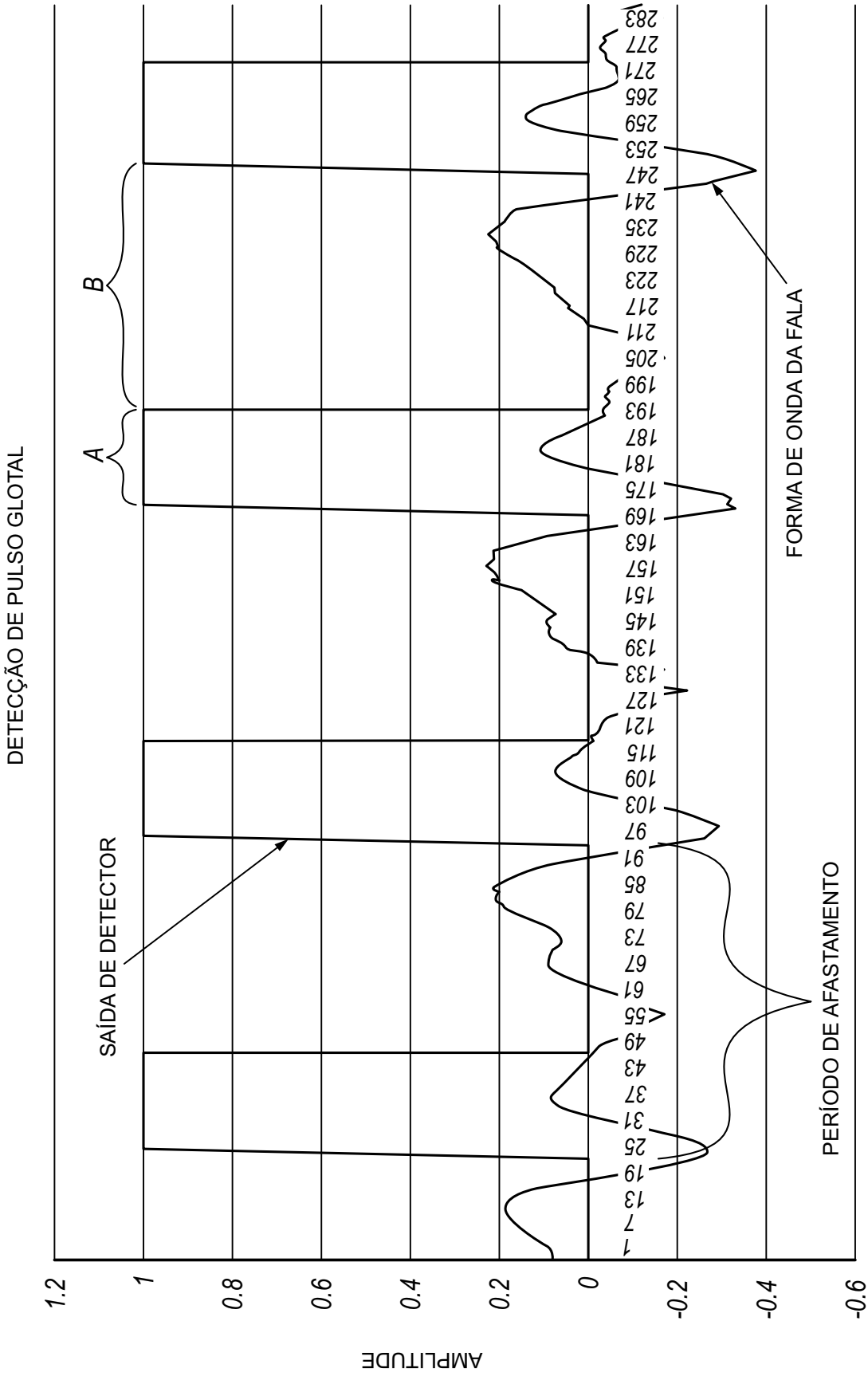
DETECÇÃO DE PULSO GLOTAL
SAÍDA DE DETECTOR
A
B
FORMA DE ONDA DA FALA
PERÍODO DE AFASTAMENTO
AMPLITUDE
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
1
7
13
19
25
31
37
43
49
55
61
67
73
79
85
91
97
103
109
115
121
127
133
139
145
151
157
163
169
175
181
187
193
199
205
211
217
223
229
235
241
247
253
259
265
271
277
283

*FIG. 14*

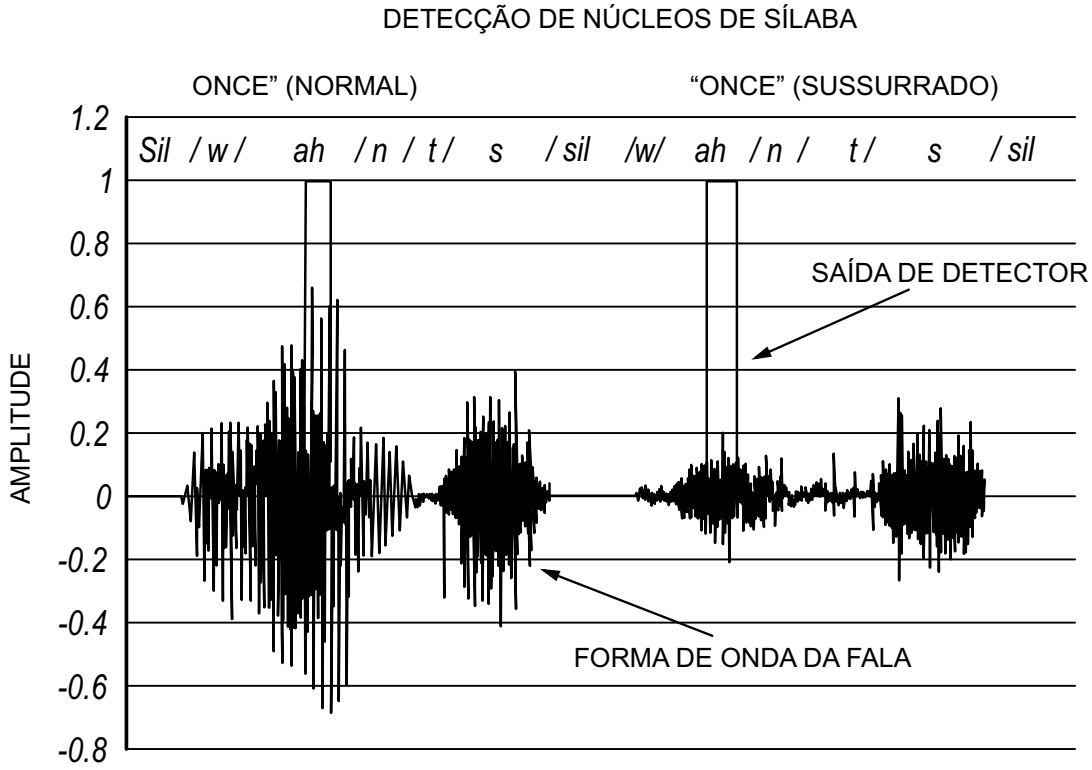
DETECÇÃO DE NÚCLEOS DE SÍLABA
ONCE" (NORMAL)
"ONCE" (SUSSURRADO)
Sil /w/ ah /n/ t/ s /sil /w/ ah /n/ t/ s /sil
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
AMPLITUDE
SAÍDA DE DETECTOR
FORMA DE ONDA DA FALA

***FIG. 15***

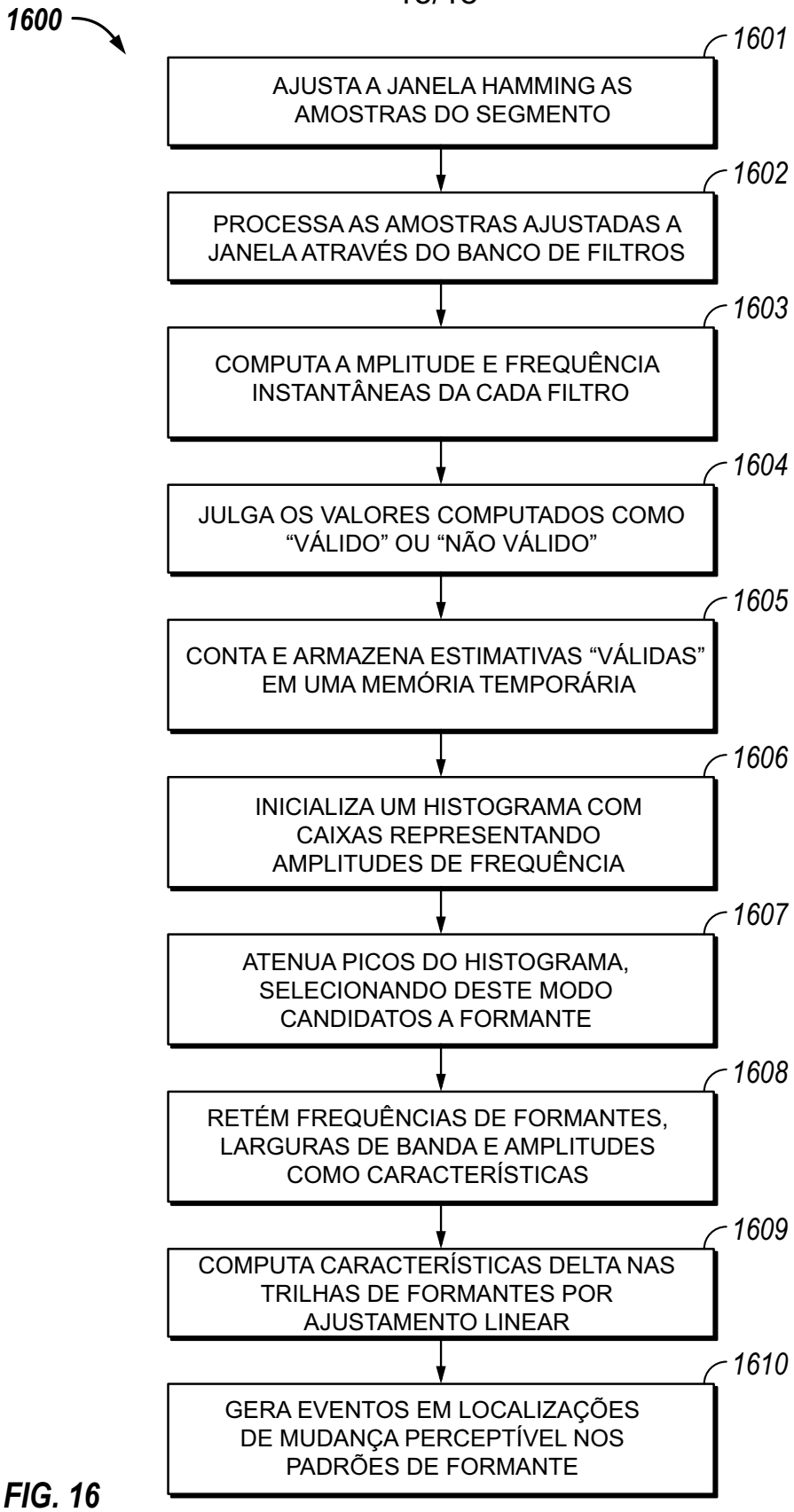
1600
1601
AJUSTA A JANELA HAMMING AS AMOSTRAS DO SEGMENTO
1602
PROCESSA AS AMOSTRAS AJUSTADAS A JANELA ATRAVÉS DO BANCO DE FILTROS
1603
COMPUTA A MPLITUDE E FREQUÊNCIA INSTANTÂNEAS DA CADA FILTRO
1604
JULGA OS VALORES COMPUTADOS COMO "VÁLIDO" OU "NÃO VÁLIDO"
1605
CONTA E ARMAZENA ESTIMATIVAS "VÁLIDAS" EM UMA MEMÓRIA TEMPORÁRIA
1606
INICIALIZA UM HISTOGRAMA COM CAIXAS REPRESENTANDO AMPLITUDES DE FREQUÊNCIA
1607
ATENUA PICOS DO HISTOGRAMA, SELECIONANDO DESTE MODO CANDIDATOS A FORMANTE
1608
RETÉM FREQUÊNCIAS DE FORMANTES, LARGURAS DE BANDA E AMPLITUDES COMO CARACTERÍSTICAS
1609
COMPUTA CARACTERÍSTICAS DELTA NAS TRILHAS DE FORMANTES POR AJUSTAMENTO LINEAR
1610
GERA EVENTOS EM LOCALIZAÇÕES DE MUDANÇA PERCEPTÍVEL NOS PADRÕES DE FORMANTE

*FIG. 16*

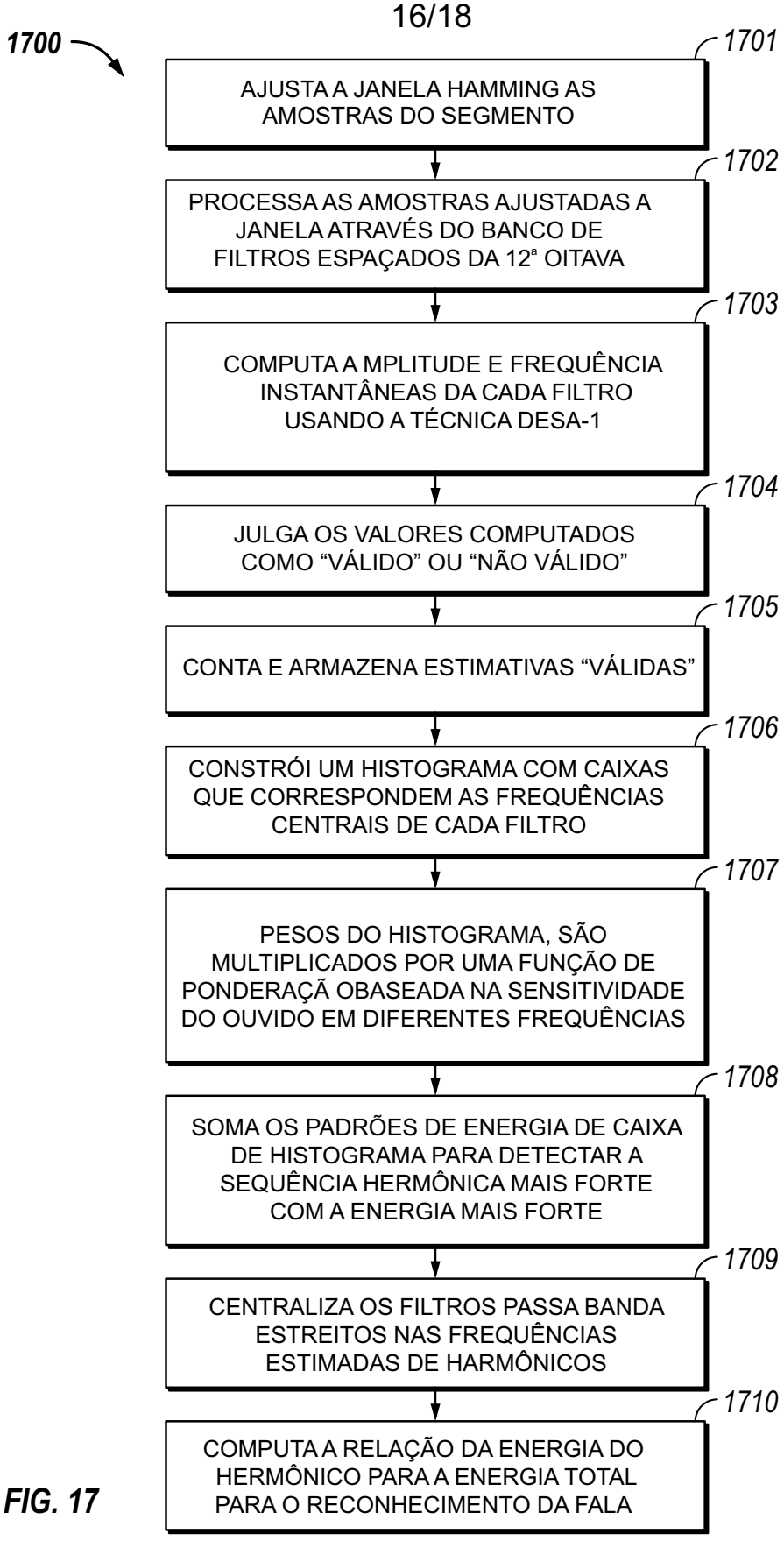
1700
1701
AJUSTA A JANELA HAMMING AS AMOSTRAS DO SEGMENTO
1702
PROCESSA AS AMOSTRAS AJUSTADAS A JANELA ATRAVÉS DO BANCO DE FILTROS ESPAÇADOS DA 12ª OITAVA
1703
COMPUTA A MPLITUDE E FREQUÊNCIA INSTANTÂNEAS DA CADA FILTRO USANDO A TÉCNICA DESA-1
1704
JULGA OS VALORES COMPUTADOS COMO “VÁLIDO” OU “NÃO VÁLIDO”
1705
CONTA E ARMAZENA ESTIMATIVAS “VÁLIDAS”
1706
CONSTRÓI UM HISTOGRAMA COM CAIXAS QUE CORRESPONDEM AS FREQUÊNCIAS CENTRAIS DE CADA FILTRO
1707
PESOS DO HISTOGRAMA, SÃO MULTIPLICADOS POR UMA FUNÇÃO DE PONDERAÇÃ OBASEADA NA SENSITIVIDADE DO OUVIDO EM DIFERENTES FREQUÊNCIAS
1708
SOMA OS PADRÕES DE ENERGIA DE CAIXA DE HISTOGRAMA PARA DETECTAR A SEQUÊNCIA HERMÔNICA MAIS FORTE COM A ENERGIA MAIS FORTE
1709
CENTRALIZA OS FILTROS PASSA BANDA ESTREITOS NAS FREQUÊNCIAS ESTIMADAS DE HARMÔNICOS
1710
COMPUTA A RELAÇÃO DA ENERGIA DO HERMÔNICO PARA A ENERGIA TOTAL PARA O RECONHECIMENTO DA FALA

*FIG. 17*

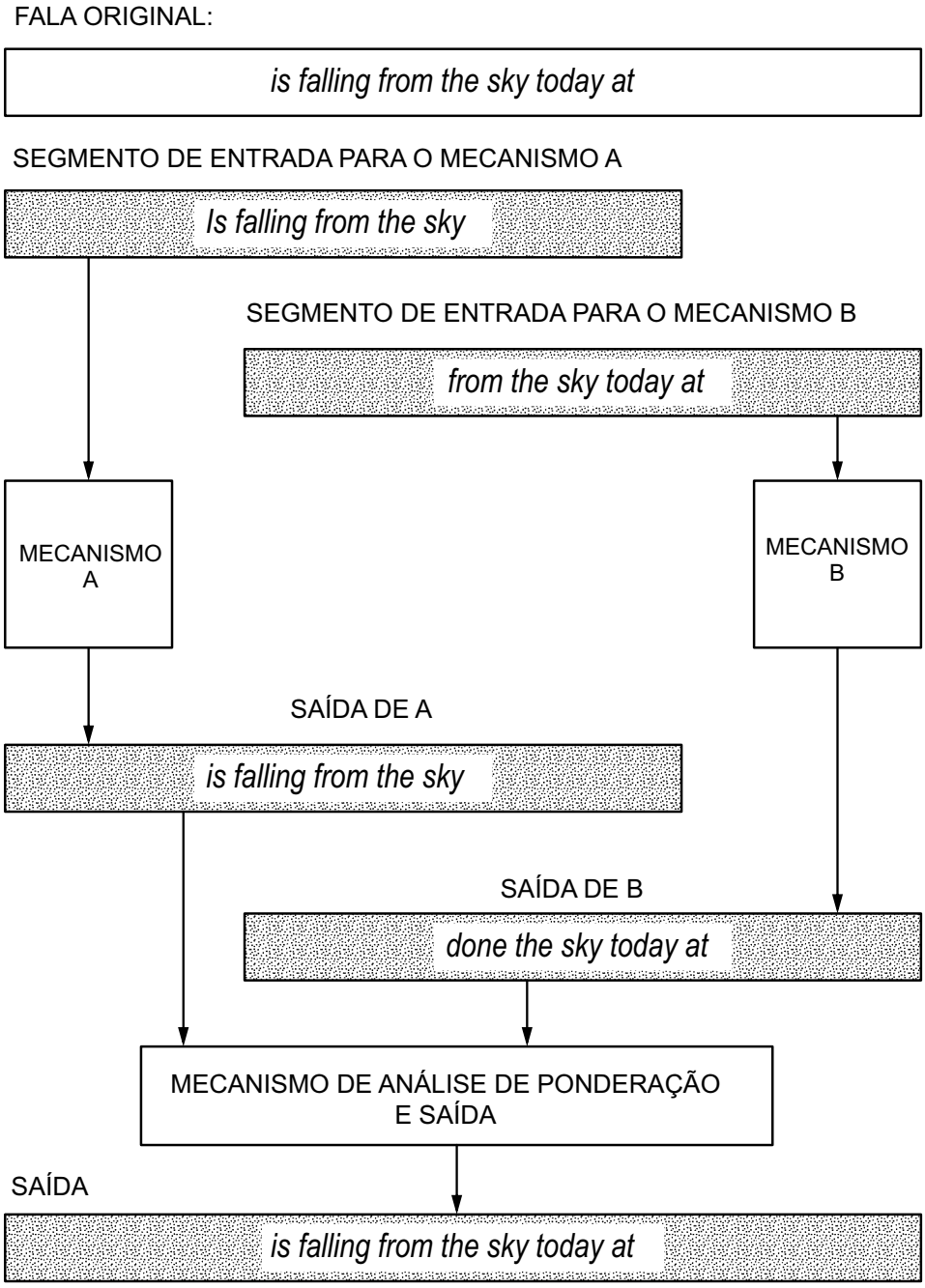
FALA ORIGINAL:
is falling from the sky today at
SEGMENTO DE ENTRADA PARA O MECANISMO A
Is falling from the sky
SEGMENTO DE ENTRADA PARA O MECANISMO B
from the sky today at
MECANISMO A
MECANISMO B
SAÍDA DE A
is falling from the sky
SAÍDA DE B
done the sky today at
MECANISMO DE ANÁLISE DE PONDERAÇÃO E SAÍDA
SAÍDA
is falling from the sky today at

***FIG. 18***

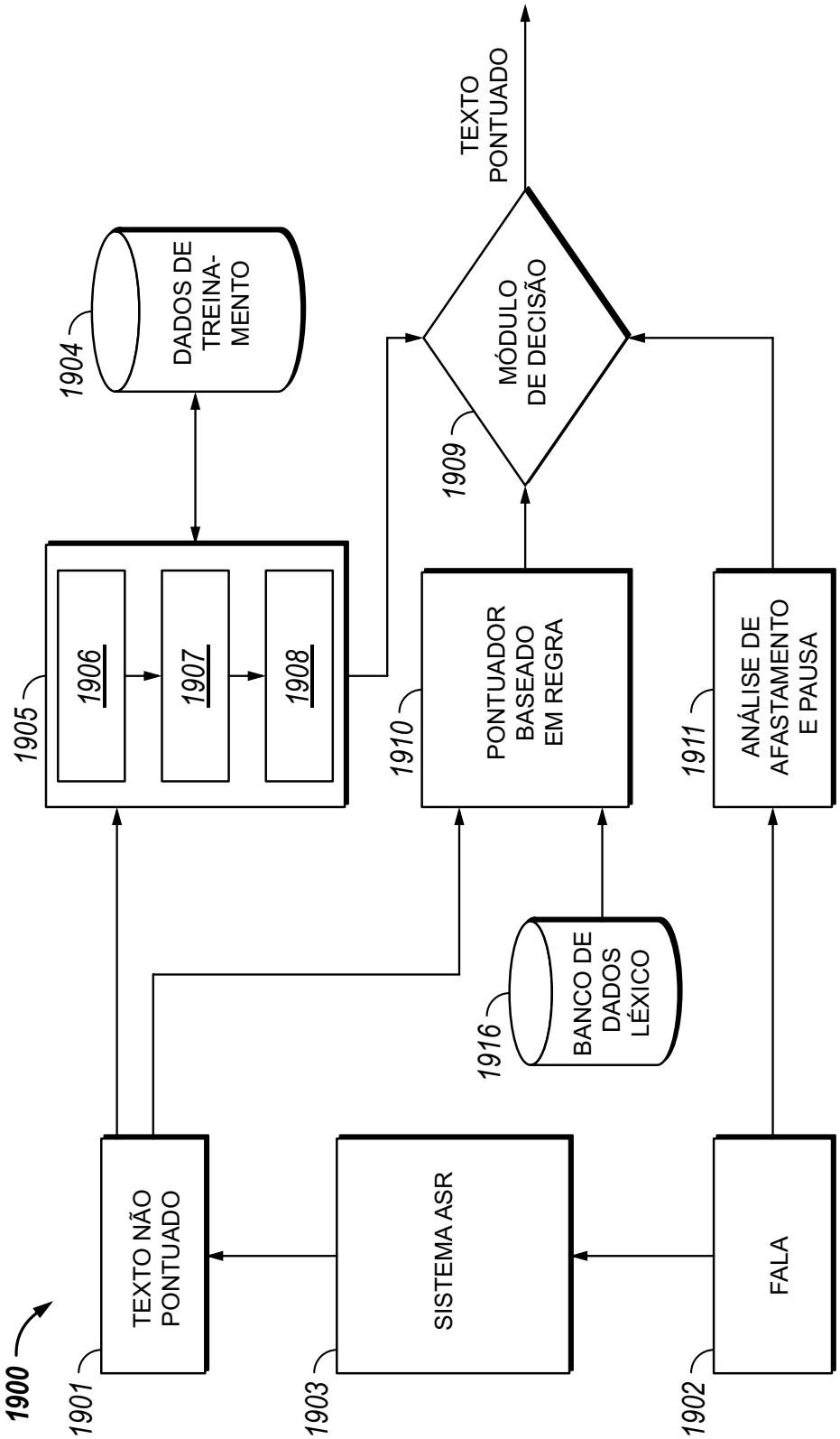
1900
1901
TEXTO NÃO PONTUADO
1903
SISTEMA ASR
1902
FALA
1916
BANCO DE DADOS LÉXICO
1905
1906
1907
1908
1904
DADOS DE TREINA-MENTO
1910
PONTUADOR BASEADO EM REGRA
1911
ANÁLISE DE AFASTAMENTO E PAUSA
1909
MÓDULO DE DECISÃO
TEXTO PONTUADO

*FIG. 19*