República Federativa do Brasil

Ministério do Desenvolvimento, Indústria, Comércio e Serviços

Instituto Nacional da Propriedade Industrial

(11) BR 112015017106-0 B1

(22) Data do Depósito: 20/07/2012

(45) Data de Concessão: 12/12/2023

(54) Título: MÉTODO IMPLEMENTADO POR COMPUTADOR PARA DETECTAR PALAVRAS-CHAVE PREDETERMINADAS

(51) Int.Cl.: G10L 15/04.

(73) Titular(es): INTERACTIVE INTELLIGENCE, INC..

(72) Inventor(es): ARAVIND GANAPATHIRAJU; ANANTH NAGARAJA IYER.

(86) Pedido PCT: PCT US2012047715 de 20/07/2012

(87) Publicação PCT: WO 2014/014478 de 23/01/2014

(85) Data do Início da Fase Nacional: 16/07/2015

(57) Resumo: MÉTODO IMPLEMENTADO POR COMPUTADOR PARA DETECTAR PALAVRAS-CHAVE PREDETERMINADAS EM UM FLUXO DE ÁUDIO E SISTEMA PARA DETECTAR PALAVRAS-CHAVE PREDETERMINADAS EM UM FLUXO DE ÁUDIO. Um sistema e método são apresentados para a análise de discurso em tempo real no campo de análise de discurso. Áudio em tempo real é alimentado juntamente com um modelo de palavra-chave, em um mecanismo de reconhecimento. O mecanismo de reconhecimento computa a probabilidade dos dados de fluxo de áudio, fazendo a compatibilidade entre palavras-chave no modelo de palavra-chave. A probabilidade é comparada a um limite em que o sistema determina se a probabilidade indica se a palavra-chave foi detectada ou não. Métricas empíricas são computadas e quaisquer alarmes falsos são identificados e rejeitados. A palavra-chave pode ser relatada como encontrada quando é estabelecido que ela não é um alarme falso e passa o limite para a detecção.

MÉTODO IMPLEMENTADO POR COMPUTADOR PARA DETECTAR PALAVRAS-CHAVE PREDETERMINADAS

RELATÓRIO

[0001] A presente invenção refere-se geralmente a sistemas e métodos de telecomunicações, bem como sistemas de reconhecimento de fala automáticos. Mais particularmente, a presente invenção refere-se à detecção de palavra-chave dentro de sistemas de reconhecimento de fala automáticos.

[0002] Sistemas de detecção de palavra-chave que estão atualmente em uso podem incluir: pesquisa fonética, modelos de lixo e reconhecimento de fala contínua de grande vocabulário (LVCSR). Cada um desses sistemas tem desvantagens inerentes que afetam a precisão e o desempenho do sistema.

[0003] Em sistemas de busca fonética, confia-se num decodificador "fonético" que converte um fluxo de áudio em uma ou várias possíveis sequências de fonemas que podem ser usadas para identificar palavras. "John diz", por exemplo, pode ser dividido na sequência de fonemas "jh aa n s eh s". O decodificador fonético faz a hipótese de um fluxo de fonema para o áudio. Esta sequência de fonema é comparada com a sequência de fonema esperada para uma palavra-chave e uma compatibilidade é encontrada. Alguns sistemas desenvolvidos com este conceito mostraram desempenho razoável, no entanto, existem muitas desvantagens para uso em uma aplicação em tempo real. Uso de um decodificador fonético antes da busca de palavra-chave, claramente, precisa ser feito em duas etapas. Isso adiciona uma complexidade considerável. Esse sistema funcionaria bem na recuperação de áudio armazenado, em que o processamento em tempo real não é necessário. Outra desvantagem é a taxa de erro com o reconhecimento de fonema. Os

identificadores de discurso no estado da técnica, que incorporam modelos de linguagem complexos, ainda produzem exatidões na faixa de 70-80%. A precisão diminui ainda mais para o discurso conversacional. Esses erros são ainda mais agravados pelos erros de busca de fonética, produzindo degradação na precisão de detecção de palavra-chave.

[0004] Outra técnica comum usada para a detecção de palavra-chave é através do uso de modelos de lixo que fazem compatibilidade com o áudio de quaisquer dados que não seja a palavra-chave. Uma rede de fonema é comumente usada para decodificar o áudio não-palavra-chave em uma sequência de fonemas. Uma abordagem simples para implementar este método é usar identificadores de discurso, em conformidade com a especificação de gramática de reconhecimento de discurso (SRGS) e escrever uma gramática da seguinte forma:

[0005] $root = $LIXO (“palavra-chave1" | “palavra-chave2”) $LIXO;

[0006] Já que a maioria dos identificadores de discurso usam decodificação fonética para implementar uma regra $LIXO, esses métodos têm as mesmas desvantagens da pesquisa fonética, especialmente do ponto de vista de uso de recursos. Outra abordagem para implementação de um modelo de lixo é tratá-lo como um estado lógico de modelo oculto de Markov (HMM) e sua probabilidade de emissão para ser uma função de todos os modelos trifone no modelo acústico, ou estimá-lo iterativamente. Ambas as abordagens prejudicavam requisitos em tempo real já que eles precisam de computação de um grande número de probabilidades ou passam pelos dados em várias passagens.

[0007] Sistemas LVCSR dependem completamente de um mecanismo de reconhecimento de discurso LVCSR para prover uma transcrição a nível de palavra do áudio e depois executar uma pesquisa de texto com base nas transcrições para a palavra-chave. Considerando o alto custo computacional de motores LVCSR, esta solução é claramente

inviável para detectar a palavra-chave em tempo real. Além disso, a precisão dos sistemas LVCSR geralmente está ligada intimamente com o conhecimento de domínio. O Vocabulário do sistema precisa ser rico o suficiente para conter todas as possíveis palavras-chave de interesse ou ser de domínio muito específico. Detectar palavras-chave de vários idiomas significaria executar múltiplos identificadores em paralelo. Um meio mais eficaz para aumentar a eficácia destes métodos é desejado para tornar detectores de palavra-chave mais difundidos em sistemas de análise de discurso em tempo real.

RESUMO

[0008] Um sistema e método são apresentados para a análise de discurso em tempo real no campo de análise de discurso. Áudio em tempo real é alimentado juntamente com um modelo de palavra-chave, em um mecanismo de reconhecimento. O mecanismo de reconhecimento computa a probabilidade dos dados de fluxo de áudio, fazendo a compatibilidade entre palavras-chave no modelo de palavra-chave. A probabilidade é comparada a um limite em que o sistema determina se a probabilidade indica se a palavra-chave foi detectada ou não. Métricas empíricas são computadas e quaisquer alarmes falsos são identificados e rejeitados. A palavra-chave pode ser relatada como encontrada quando é estabelecido que ela não é um alarme falso e passa o limite para a detecção.

[0009] Em uma modalidade, é divulgado um método implementado em computador para detectar palavras-chave predeterminadas em um fluxo de áudio, compreendendo as etapas de: a) desenvolver um modelo de palavra-chave para as palavras-chave predeterminadas; b) comparar o modelo de palavra-chave e o fluxo de áudio para detectar as prováveis das palavras chave predeterminadas; c) computar uma probabilidade de que uma porção do fluxo de áudio é compatível a uma das palavras chave predeterminadas do modelo de

palavra-chave; d) comparar a probabilidade calculada a um limite predeterminado; e) declarar uma palavra detectada em potencial se a probabilidade computada for maior que o limite predeterminado; f) computar mais dados para auxiliar na determinação de incompatibilidades; g) usar os dados adicionais para determinar se a palavra detectada em potencial é um alarme falso; e h) relatar a palavra-chave detectada se um alarme falso não for identificado na etapa (g).

[00010] Em outra modalidade, é divulgado um método implementado em computador para detectar palavras-chave predeterminadas em um fluxo de áudio, compreendendo as etapas de: a) desenvolver um modelo de palavra-chave para as palavras-chave predeterminadas; b) dividir o fluxo de áudio em uma série de pontos em um espaço acústico que abrange todos os sons possíveis criados em uma linguagem específica; c) computar uma probabilidade posterior que uma primeira trajetória de cada modelo de palavra-chave para as palavras-chaves predeterminadas no espaço acústico seja compatível com uma segunda trajetória de uma porção de uma série de pontos no espaço acústico; d) comparar a probabilidade posterior a um limite predeterminado; e e) relatar um palavra-chave detectada, se a probabilidade posterior for maior que o limite predeterminado.

[00011] Em outra modalidade, é divulgado um sistema implementado em computador para detectar palavras-chave predeterminadas em um fluxo de áudio, compreendendo: meios de desenvolver um modelo de palavra-chave para as palavras-chave predeterminadas; meios de comparar o modelo de palavra-chave e o fluxo de áudio para detectar as prováveis das palavras chave predeterminadas; meios de computar uma probabilidade de que uma porção do fluxo de áudio é compatível a uma das palavras chave predeterminadas do modelo de palavra-chave; meios de comparar a probabilidade calculada a um limite predeterminado; meios de declarar uma palavra detectada em potencial se

a probabilidade computada for maior que o limite predeterminado; meios de computar mais dados para auxiliar na determinação de incompatibilidades; meios de usar os dados adicionais para determinar se a palavra detectada em potencial é um alarme falso; e meios de relatar a palavra-chave detectada se um alarme falso não for identificado.

BREVE DESCRIÇÃO DAS FIGURAS

[00012] A Figura 1 é um diagrama que ilustra os componentes básicos de um detector de palavra-chave.

[00013] A Figura 2 é um diagrama ilustrando um modelo HMM concatenado.

[00014] A Figura 3a é um diagrama ilustrando uma visualização abstrata do espaço de recurso de áudio e os modelos de trifone que abrangem este espaço.

[00015] A Figura 3b é um diagrama ilustrando modelos monofone que abrangem completamente o mesmo espaço de recurso de áudio.

[00016] A Figura 4 é um diagrama ilustrando um sinal de discurso, mostrando uma palavra-chave falada rodeada por modelos lixo.

[00017] A Figura 5 é uma tabela ilustrando as probabilidades de nível de fonema.

[00018] Figura 6 é um diagrama ilustrando a relação entre a compatibilidade interna "Pontuação" e valores externos de "Certeza".

[00019] A Figura 7 é um diagrama ilustrando o comportamento do sistema com configurações variadas de certeza.

[00020] Figura 8 é um fluxograma ilustrando o algoritmo para detectar palavra-chave utilizado no sistema.

DESCRIÇÃO DETALHADA

[00021] Com a finalidade de promover uma compreensão dos princípios da invenção, vai ser feita agora referência às modalidades ilustradas nas figuras e linguagem específica será usada para descrever as mesmas. Será, no entanto, compreendido que nenhuma limitação do

escopo da invenção é assim pretendida. Quaisquer alterações e modificações adicionais nas modalidades descritas e quaisquer outras aplicações dos princípios da invenção conforme descrito neste documento são contempladas, como normalmente ocorreria a um versado na técnica a qual se refere a invenção.

[00022] Sistemas de reconhecimento automático de discurso (ASR) analisam o discurso humano e traduzem-no em texto ou palavras. O desempenho destes sistemas é comumente avaliado com base na exatidão, confiabilidade, suporte de idioma e a velocidade com a qual o discurso pode ser reconhecido. O desempenho do sistema deve ser muito alto. Desempenho superior, muitas vezes é quantificado por uma taxa de detecção alta e uma taxa baixa de alarme falso. O padrão da indústria é considerado sendo em torno de uma taxa de detecção de 70% em 5 alarmes falsos por palavra-chave por hora de discurso, ou 5 FA/kw/hr. Fatores tais como sotaque, articulação, taxa de discurso, pronúncia, ruído de fundo, etc., podem ter um efeito negativo sobre a precisão do sistema. A velocidade de processamento é necessária para analisar várias centenas de conversas telefônicas ao mesmo tempo e em tempo real. Também espera-se do sistema que ele execute consistentemente e confiantemente, independentemente de condições de canal e vários artefatos introduzidos por canais de telefonia modernos, especialmente voz sobre IP. Palavras-chave de vários idiomas também precisam ser detectadas na mesma fonte de áudio.

[00023] Aqueles versados na técnica vão reconhecer da divulgação presente que as várias metodologias divulgadas neste documento podem ser implementadas em computador usando muitas formas diferentes de equipamentos de processamento de dados, tais como microprocessadores digitas e memória associada executando programas de software apropriados, para citar apenas um exemplo de não-limitação. A forma específica do hardware, firmware e software usados para

implementar as modalidades atualmente divulgadas não é crítica para a presente invenção.

[00024] Na presente invenção, computações de probabilidade posterior para sistemas de reconhecimento de discurso podem ser usadas para aumentar a eficácia do sistema. Sistemas anteriores projetados para executar detecção de palavra-chave usam a medida da probabilidade de log para compatibilizar o áudio apresentado aos fonemas em uma palavra-chave. Fonemas são unidades sub palavra que normalmente são modelados em sistemas ASR. Além disso, os fonemas podem ser modelados isoladamente ou no contexto de outros fonemas. Os primeiros são chamados de monofones e os últimos são chamados de trifones quando o fonema depende de seu contexto fonêmico anterior e próximo. Probabilidade posterior, como usada nesta invenção, pode ser uma medida de quão bem o áudio é compatível a um modelo quando comparado com o mesmo áudio enquanto é feita a compatibilidade a todos os outros modelos para um determinado padrão de discurso.

[00025] O uso de probabilidades posteriores no reconhecimento de discurso foi tentado no passado, principalmente pelo treinamento de uma rede neural. Enquanto este método retorna uma aproximação à probabilidade posterior, ele tende a ser extremamente computadoramente dispendioso e requer procedimentos de treinamento especial.

[00026] Uma abordagem alternativa ao cálculo de probabilidade posterior para reconhecimento de discurso pode ser desenvolvida conforme segue:

[00027] Por definição, a probabilidade posterior (P) de um modelo (T_i), dado um vetor de observação x, pode ser escrita como:

[00028] $$P(T_i \mid x) = \frac{P(x \mid T_i)P(T_i)}{\sum_j P(x \mid T_j)P(T_j)}$$

[00029] Em que P ($x|T_i$) é a probabilidade do modelo T_i gerar a acústica x e j é uma variável que abrange os índices de todos os modelos.

Na equação acima, o termo P (T_i) é mantido constante para todos os modelos e a fórmula pode ser reescrita como:

[00030] $$P(T_i \mid x) = \frac{P(x \mid T_i)}{\sum_j P(x \mid T_j)}$$

[00031] Esta equação ainda é proibitivamente dispendiosa para calcular. A despesa pode ser atribuída ao fato de que o termo do denominador é um somatório de todos os modelos, que pode ser muito grande para um sistema com base em trifone dependente de contexto (tipicamente dezenas de milhares de modelos). Para estudar o impacto dos termos denominadores, pode-se considerar uma abordagem gráfica e intuitiva. O denominador como um todo significa a probabilidade total de modelos abrangendo todo o espaço de áudio. Portanto, a equação acima pode ser reescrita como:

[00032] $$P(T_i \mid x) = \frac{P(x \mid T_i)}{\sum \mathcal{M}; \forall \mathcal{M} \in \mathbb{M}}$$

[00033] Em que M representa um modelo, ∀M representa todos os modelos em todo o espaço de áudio, representados como M.

[00034] A fórmula acima não perde a generalidade. O termo denominador é agora um somatório sobre qualquer conjunto de modelos que abrange completamente o espaço de recurso de áudio.

[00035] A Figura 1 é um diagrama que ilustra os componentes básicos de um detector de palavra-chave, 100. Os componentes básicos de um detector de palavra-chave 100 podem incluir dados/palavras-chave de usuário 105, modelo de palavra-chave 110, fontes de conhecimento 115 que incluem um modelo acústico 120 e um Preditor/dicionário de pronúncia 125, um fluxo de áudio 130, uma calculadora de recurso front-end 135, um mecanismo de reconhecimento (compatibilidade por padrão) 140 e o relato de palavras-chave encontradas em tempo real 145.

[00036] Palavras-chave podem ser definida, 105, pelo usuário do sistema de acordo com a preferência do usuário. O modelo de palavra-

chave 110 pode ser formado pela concatenação de fonema HMMs. Isto é descrito adicionalmente na descrição da Figura 2. O modelo de palavra-chave, 110, pode ser composto com base em palavras-chave que são definidas pelo usuário e a entrada para o modelo de palavra-chave com base em fontes de conhecimento, 115. Tais fontes de conhecimento podem incluir um modelo acústico, 120 e um preditor/dicionário de pronúncia, 125.

[00037] As fontes de conhecimento 115 podem armazenar modelos probabilísticos das relações entre os eventos acústicos e pronúncias. As fontes de conhecimento 115 podem ser desenvolvidas através da análise de grandes quantidades de dados de áudio. O modelo acústico e o preditor/dicionário de pronúncia são feitos, por exemplo, olhando para uma palavra como "hello" e examinando os fonemas que compõem a palavra. Cada palavra-chave no sistema é representada por um modelo estatístico de suas unidades de sub-palavra constituintes chamado os fonemas. Os fonemas de "hello", conforme definido em um dicionário de fonema padrão são: "hh", "eh", "l" e "ow". Modelos dos quatro fonemas são então amarrados combinados juntos em um modelo composto que então se torna o modelo de palavra-chave para a palavra "hello". Estes modelos são dependentes de idioma. Para também oferecer suporte a vários idiomas, podem prever-se várias fontes de conhecimento.

[00038] O modelo acústico 120 pode ser formado por modelagem estatística dos vários sons que ocorrem em um determinado idioma. Um fonema é considerado a unidade básica de som. Presume-se um conjunto predefinido de tais fonemas para descrever completamente todos os sons de uma língua particular. Um HMM, que codifica a relação entre o sinal de áudio observado e os fonemas não observados, forma a teoria fundamental para a maioria dos sistemas de reconhecimento de discurso modernos. Um fonema é considerado composto de três estados, que representam as porções iniciais, centrais e posterior do som. Um HMM é construído pela concatenação desses três estados. Um processo de formação estuda as

propriedades estatísticas de cada um desses estados para todos os fonemas sobre uma grande coleção de áudio transcrito. Uma relação entre as propriedades textuais e as propriedades faladas é então formada. Normalmente, as estatísticas dos estados podem ser codificadas usando um modelo de mistura gaussiano (GMM). Um conjunto destes GMMs é denominado como um modelo acústico. Especificamente, o descrito nesta aplicação é referido como um modelo monofone, ou independente de contexto. Muitos outros tipos de modelo também podem ser usados. Por exemplo, muitos sistemas de reconhecimento de discurso modernos podem utilizar um modelo acústico mais avançado, que pode ser dependente de contexto e capturar as variações complexas criadas devido à posição dos fonemas no discurso conversacional. Cada estado de um fonema é especializado para seus fonemas vizinhos esquerdo e direito. Claramente, tal esquema resultaria em um número muito grande de GMMs no modelo acústico. Um exemplo de um fonema dependente de contexto é um trifone.

[00039] O dicionário de pronúncia, 125, na Figura 1 pode ser responsável pela decomposição de uma palavra em uma sequência de fonemas. Palavras-chave apresentadas do usuário pode ser na forma legível por humanos, como grafemas/alfabetos de uma linguagem específica. No entanto, o algoritmo de compatibilidade por padrão pode contar com uma sequência de fonemas que representam a pronúncia da palavra. A presente invenção utiliza um dicionário de pronúncia, que pode armazenar um mapeamento entre palavras comumente faladas e suas pronúncias. Uma vez obtida a sequência de fonemas, o modelo estatístico correspondente para cada um dos fonemas no modelo acústico pode ser examinado. Uma concatenação destes modelos estatísticos pode ser usada para executar a detecção de palavra-chave para a palavra de interesse. Para palavras que não estão presentes no dicionário, um preditor, que é baseado em regras linguísticas, pode ser usado para resolver as pronúncias.

[00040] O fluxo de áudio (ou seja, o que é falado no sistema pelo usuário), 130, pode ser alimentado na calculadora de recurso front-end, 135, que pode converter o fluxo de áudio em uma representação do fluxo de áudio, ou uma sequência de características espectrais. Análise de áudio pode ser realizada, ao se segmentar o sinal de áudio como uma sequência de janelas curtas (tipicamente de 10 ms) e extrair recursos de domínio espectral. Para cada janela, a calculadora de recurso pode calcular um conjunto de 13 coeficientes cepstrais da frequência mel (MFCC) e seus derivados de primeira e segunda ordem. Os cálculos resultantes representam cada uma destas janelas como um ponto em um espaço tridimensional 39 M. Este espaço abrange completamente todos os sons possíveis, criados em um determinado idioma.

[00041] O modelo de palavra-chave, 110, que pode ser formado pela concatenação de fonemas de modelos ocultos de Markov (HMMs) e o sinal do fluxo de áudio, 135, ambos podem então ser alimentados em um mecanismo de reconhecimento de compatibilidade por padrão, 140. A tarefa do mecanismo de reconhecimento pode ser tomar um conjunto de modelos de palavra-chave e pesquisa através de fluxo de áudio apresentado para determinar se as palavras foram faladas. No espaço multidimensional construído pela calculadora de recurso, uma palavra falada pode tornar-se uma sequência de vetores MFCC formando uma trajetória no espaço acústico M. Detecção de palavra-chave pode agora simplesmente tornar-se um problema de probabilidade de computação de gerar a trajetória de acordo com o dado modelo de palavra-chave. Esta operação pode ser alcançada usando o princípio bem conhecido de programação dinâmica, especificamente o algoritmo de Viterbi, que alinha o modelo de palavra-chave para o melhor segmento do sinal de áudio e resulta em uma pontuação de compatibilidade. Se a pontuação de compatibilidade é significativa, o algoritmo de detecção de palavra-chave

infere que a palavra-chave foi falada e relata um evento de palavra-chave detectada.

[00042] As palavras-chave resultantes podem ser relatadas, então, em tempo real, 145. O relatório pode ser apresentado como um tempo de início e de fim da palavra-chave no fluxo de áudio com um valor de certeza que a palavra-chave foi encontrada. O valor primário de certeza pode ser uma função de como a palavra-chave é falada. Por exemplo, no caso de várias pronúncias de uma única palavra, a palavra-chave "tomato" pode ser falado como "te-mah-toh" e "te-may-toh". O valor primário de certeza pode ser menor quando a palavra é falada em uma pronúncia menos comum ou quando a palavra não é bem enunciada. A variante específica da pronúncia que faz parte de um reconhecimento específico também é exibida no relatório.

[00043] A Figura 2 é um diagrama ilustrando um modelo HMM concatenado. Um modelo de palavra-chave pode ser formado pela concatenação de fonema HMMs. Por exemplo, o modelo de palavra-chave 200 para a palavra "rise" é construído a partir dos modelos monofone dos fonemas que compõem a sua pronúncia. Os fonemas que compreendem a pronúncia de "rise" são "r", "ay" e "z". Cada fonema tem três estados presentes consistindo de uma porção de início do som 210, uma porção central de som 211 e porção posterior de som 212. Por exemplo, o fonema "r" tem uma porção de início de som 210 mostrada como "r1" no modelo. A porção central de som 211 é exibida por "r2" e a porção posterior de som 212 é exibida por "r3". O fonema "Ay" tem uma porção de início de som 210 ilustrada como "ay1" no modelo. A porção central de som 211 é ilustrada por "ay2" e a porção posterior de som 212 é ilustrada por "ay3". O fonema "z" tem uma porção de início de som 210 ilustrada como "z" no modelo. A porção central de som 211 é exibida por "z2" e a porção posterior de som 212 é exibida por "z3". Cada porção de som tem uma transição 213 dentro da porção em si ou entre porções. De forma semelhante, um modelo de

palavra-chave dependente de contexto pode ser construído pela concatenação de seus modelos trifone.

[00044] A Figura 3a é um diagrama ilustrando uma visualização abstrata do espaço de recurso de áudio e os modelos de trifone que abrangem este espaço. Na realidade, o espaço de áudio é 39-dimensional, mas para fins de ilustração, é mostrado um espaço 2-dimensional. A Figura 3b é um diagrama ilustrando modelos monofone que abrangem completamente o mesmo espaço de recurso de áudio. Tendo em conta as observações das figuras 3a e 3b, o algoritmo de palavra-chave como apresentado acima

[00045] $$P(T_i \mid x) = \frac{P(x \mid T_i)}{\sum \mathcal{M}; \forall \mathcal{M} \in \mathbb{M}}$$

[00046] Torna-se

[00047] $$P(T_i \mid x) = \frac{P(x \mid T_i)}{\sum_k P(x \mid M_k)}$$

[00048] Quando M infere-se como o conjunto de modelos de monofone na primeira equação, e em que M_k representa os modelos monofone na segunda equação. ∀M infere-se como o conjunto de modelos monofone. Será apreciado da presente divulgação que T_i e M_k ambos abrangem o espaço inteiro de áudio, M, completamente. Já que o número de GMMs presente no modelo monofone (Figura 3b) é significativamente menor comparado ao modelo trifone (Figura 3a),a computação de probabilidades posteriores é extremamente rápida, mas uma representação próxima do valor correto.

[00049] A Figura 4 é um diagrama ilustrando um sinal de discurso 400, mostrando uma palavra-chave falada 410 rodeada por modelos lixo 405, 415. Uma palavra-chave é falada como parte de um fluxo contínuo de discurso. No segmento de áudio entre t_o e t_s, o modelo de lixo 405 toma precedência, enquanto faz a compatibilidade com porções de áudio não-palavra-chave. A pontuação acumulada durante este período é

representada por S_1 nas seguintes equações. Da mesma forma, no segmento de áudio t_e a t_N, a pontuação de compatibilidade de lixo é representada por S_2. Aqui, o modelo de lixo 415 toma precedência. Em vez de explicitamente computar as probabilidades de lixo, de S_1 e S_2, um valor constante e é escolhido tal que

[00050] $$e(T_s - T_0) = S_1,$$

[00051] E

[00052] $$e(T_N - T_e) = S_2.$$

[00053] A constante e é validada em um grande conjunto de dados do teste para não realizar qualquer redução significativa no desempenho quando comparada à computação explicita da probabilidade de lixo. Esta aproximação de usar um valor de lixo constante torna o sistema significativamente mais rápido em comparação a algoritmos de detecção de palavra-chave tradicionais.

[00054] Figura 5 é uma tabela ilustrando as probabilidades de nível de fonema 500 comparando as probabilidades de compatibilidade de fonema das palavras faladas "December" e "Discover" em comparação com o modelo de palavra-chave para "December". Uma taxa elevada de alarmes falsos pode ser contada como um dos principais problemas em um algoritmo de detecção de palavra-chave. Ao contrário de mecanismos LVCSR, detectores de palavra-chave não têm acesso a informação contextual a nível de palavra. Por exemplo, ao procurar a palavra-chave "rise", o sinal sonoro de "rise" é muito semelhante ao de "price", "rice", "prize", "notarize", etc. Desta maneira, estas palavras, iriam ser tratadas como uma compatibilidade pelo sistema. Este é um problema semelhante ao de pesquisas de subsequência de caracteres no texto onde subpalavras são compatíveis com a sequência chave.

[00055] A fim de conter alarmes falsos, segue alguns exemplos não limitantes de abordagens que podem ser usadas como uma verificação secundária de compatibilidade de palavras-chave encontradas pelo

algoritmo de Viterbi principal. Anti-palavras são um conjunto de palavras que são comumente confundidas com palavras-chave dentro do sistema. No exemplo apresentado com as palavras “price”, “rice”, “prize”,“notarize”, etc., como mencionado acima, estas palavras compõem o conjunto de anti-palavra da palavra-chave "rise". O sistema procura estas anti-palavras em paralelo com a palavra-chave e relata um evento de palavra-chave encontrada apenas quando a pontuação de compatibilidade de palavra-chave supera a pontuação de compatibilidade de anti-palavra. Esse recurso é um método eficaz para reduzir espúrios de alarmes falsos. No entanto, o método ainda requer a intervenção do usuário e a criação de grandes conjuntos de anti-palavras. Outras técnicas podem ser puramente orientadas por dados e, portanto, às vezes mais desejáveis.

[00056] A percentagem de incompatibilidade de fonema determina o número de fonemas da palavra-chave que tem incompatibilidade com o sinal de áudio, mesmo que a probabilidade no geral de palavra-chave da busca Viterbi foi encontrada como uma compatibilidade. Por exemplo, a palavra "December", como mostrada na Figura 5, pode ser encontrada correspondendo erroneamente a instâncias de "Discover" pelo detector de palavra-chave. Probabilidades de nível de fonema são exemplificadas na Figura 5. A pontuação representa quanto o fonema é compatível com o fluxo de áudio. Usando o exemplo de imediato, quanto mais positivo for o número, melhor a compatibilidade. Um valor de pontuação "0" indica uma compatibilidade perfeita. Estas pontuações são sempre negativas ou zero. Para o fonema "d", a probabilidade para "December" é-0.37, enquanto é -1.18 para "discover". Pode-se notar que todos os fonemas rendem menores probabilidades quando a elocução falada foi "discover" em comparação com a elocução falada "December". Essa métrica computa a porcentagem de tais fonemas que não combinam e executa uma verificação adicional antes de relatar o evento de palavra-chave encontrada.

[00057] Análoga à percentagem de incompatibilidade de fonema, a medida de percentagem de compatibilidade de fonema computa a percentagem de fonemas que são compatíveis com o sinal de áudio. O percentual de fonemas que combinam deve ser acima de um limite pré-definido para que o evento de palavra-chave encontrada seja relatado.

[00058] A probabilidade penalizada por duração enfatiza incompatibilidades duracionais de uma palavra-chave com o fluxo de áudio. Por exemplo, consoantes como "t", "d" e "b" têm uma duração esperada inferior em comparação com as vogais como "aa", "ae" e "uw". No caso de estas consoantes serem compatíveis por mais do que a duração esperada, a correspondência de palavra-chave é, provavelmente, um alarme falso. Esses eventos podem ser o resultado de modelo acústico pobre ou presença de ruído no sinal sendo analisado. Para capturar tal cenário, a probabilidade penalizada por duração é computadorizada como

[00059] $$\tilde{p}_i = \begin{cases} 2p_i, & \text{if } d_i > D \\ p_i, & \text{if } d_i <= D \end{cases}$$

[00060] Em que pi representa a probabilidade de fonema i, di representa a duração de fonema ie D representa um limite de duração determinado com base em testes realizados em grandes conjuntos de dados. A pontuação penalizada por duração para uma palavra-chave pode ser representada pela média de todas as suas pontuações de fonema. Ao se dobrar as pontuações para fonemas longos, esta métrica enfatiza incompatibilidades criadas por fonemas espúrios, e assim, diminui alarmes falsos.

[00061] Figura 6 é um diagrama ilustrando a relação entre a compatibilidade interna "Pontuação" e valores externos de "Certeza". Índice de detecção é uma medida de precisão esperada do sistema. O uso primário desta medida é orientar os usuários na determinação de um bom conjunto de palavras-chave. Outros usos incluem feedback para o mecanismo de reconhecimento e controlar a taxa de alarme falso. O

diagrama na Figura 6 mostra a relação entre a probabilidade de correspondência, ou a "pontuação", conforme determinado pelo mecanismo de reconhecimento e os valores de certeza, conforme relatado pelo sistema. Por padrão, a curva sólida 605 é usada se nenhuma informação sobre a palavra-chave for conhecida. Se o índice de detecção for conhecida, a relação pode ser modificada, alterando-se a faixa operacional de pontuação da palavra-chave, conforme mostrado pelas linhas tracejadas e pontilhadas. A linha tracejada 610 apresenta uma palavra-chave de baixo índice de detecção, enquanto a linha pontilhada 615 exibe uma palavra-chave de alto índice de detecção. À medida que aumenta o valor de certeza, assim como a probabilidade de uma compatibilidade em que 0.0 é indicativo não compatibilidade e 1.0 é uma compatibilidade. À medida que a minPontuação torna-se mais negativa, também a probabilidade de uma incompatibilidade. À medida que a pontuação se aproxima de 0.0, há uma maior probabilidade de uma compatibilidade. Assim, uma pontuação de 0 e uma certeza de 1.0 indica uma combinação perfeita.

[00062] A Figura 7 é um diagrama ilustrando o comportamento do sistema com configurações variadas de certeza. O resultado da mudança da faixa operacional com base no índice de detecção é um comportamento mais controlado do sistema. Quando um usuário registra uma palavra-chave a ser detectada, uma medida de índice de detecção associada é apresentada, tal como 70. Por definição, isso significa que o sistema resulta em 70% de precisão com uma taxa de alarme falso de 5 por hora. Para obter esse comportamento do sistema, a faixa de pontuação interna é modificada conforme mostrado na Figura 7, tal que na configuração de confiança padrão (0,5) o sistema produz 5 alarmes falsos por hora e uma taxa de detecção de 70%. Se o usuário deseja uma precisão mais elevada, a configuração de certeza é reduzida, que por sua vez possivelmente poderia criar uma maior taxa de alarme falso. Se o usuário deseja menor taxa de alarme falso, a configuração de certeza é aumentada,

possivelmente resultando em menor taxa de detecção.

[00063] O diagrama 700 ilustra o comportamento do sistema conforme alteradas as configurações de certeza. Conforme a configuração de certeza se aproxima a 1.0, a taxa de detecção diminui até atingir um valor 0.0 em uma configuração de certeza de 1.0. A taxa de alarmes falsos também diminui e se aproxima a 0.0 conforme a configuração de certeza se aproxima a 1.0. Por outro lado, conforme a taxa de detecção aumenta, a configuração de certeza se aproxima a 0,0 e a taxa de alarmes falsos (FA/Hr) aumenta.

[00064] Conforme ilustrado na Figura 8, um processo 800 para utilizar o algoritmo de detecção de palavra-chave é provido. O processo 800 pode ser operativo em qualquer ou todos os elementos do sistema 100 (Figura 1).

[00065] Os dados ficam contidos tanto no modelo de palavra-chave 805 e o fluxo de áudio 810. Enquanto o modelo de palavra-chave 805 pode ser necessário apenas uma vez durante o processo de fluxo de dados, o fluxo de áudio 810 é uma contínua alimentação de dados no sistema. Por exemplo, o fluxo de áudio pode ser uma pessoa falando no sistema em tempo real através de um telefone digital. O modelo de palavra-chave 805, que é formado pela concatenação de fonemas HMMs, contém as palavras-chave que são definidas pelo usuário de acordo com a preferência do usuário. Por exemplo, um usuário pode definir palavras-chave que são específicas à indústria como "termos", "condições", "premium" e "apoio" para a indústria de seguros. Essas palavras-chave no modelo de palavra-chave 810 são utilizadas para compatibilidades por padrão com as palavras que estão continuamente em entrada no sistema através do fluxo de áudio 810. O controle é passado à operação 815 e o processo 800 continua.

[00066] Na operação 815, a probabilidade é computada no Mecanismo de Reconhecimento, 140 (Figura 1). Como descrito anteriormente, pontuações de probabilidade são usadas pelo sistema para determinar os fonemas compatíveis. O percentual destes fonemas deve ser acima do limite pré-definido para que o evento de palavra-chave encontrada seja relatado. O controle é passado à operação 820 e o processo 800 continua.

[00067] Na operação 820, determina-se se a probabilidade computada é maior ou não que o limite. Se for determinado que a probabilidade é maior que o limite, então o controle é passado para a etapa 825 e o processo 800 continua. Se for determinado que a probabilidade não é maior que o limite, então o controle de sistema é passado para a etapa 815 e processo 800 continua.

[00068] A determinação em operação 820 pode ser feita com base em quaisquer critérios adequados. Por exemplo, o limite pode ser definido pelo usuário ou deixado em um valor padrão do sistema. Conforme o valor do limite, ou configuração de certeza, aproxima-se de 0.0, maior a frequência de alarmes falsos que podem ocorrer. A taxa de detecção da palavra-chave não pode ser muito maior do que se a configuração de certeza for ligeiramente maior com menos frequência de alarmes falsos.

[00069] Caso o controle seja passado de volta à etapa 815, a probabilidade é então computada novamente usando um pedaço diferente do fluxo de áudio e o processo continua.

[00070] Na operação 825, o sistema calcula métricas empíricas, tais como comparação a pontuações de anti-palavras, percentagem de incompatibilidade de fonema, percentagem de compatibilidade de fonema, e/ou probabilidade penalizada por duração, para citar apenas alguns exemplos não-limitantes. As métricas são usadas para computar dados secundários e podem servir como uma verificação adicional antes de relatar

eventos de palavra-chave encontrada. O controle é passado à operação 830 e o processo 800 continua.

[00071] Na operação 830, é determinado se as possíveis compatibilidades são identificadas como alarmes falsos ou não. Se for determinado que as combinações possíveis são alarmes falsos, então o controle é passado para a etapa 815 e processo 800 continua. Se for determinado que as possíveis compatibilidades não são alarmes falsos, então o controle é passado para a etapa 835 e processo 800 continua.

[00072] Uma vez que o processo retorna para a etapa 815, probabilidade é computada novamente usando um pedaço diferente do fluxo de áudio e o processo continua.

[00073] A determinação em operação 830 pode ser feita com base em quaisquer critérios adequados. Em algumas modalidades, os critérios são baseados nas probabilidades e as métricas empíricas que foram calculadas pelo sistema.

[00074] Na operação 835, o sistema informa a palavra-chave como encontrada e o processo termina.

[00075] Embora a invenção tenha sido ilustrada e descrita em detalhes nas figuras e descrição acima, a mesma é para ser considerada como ilustrativa e não restritiva em caráter, subentendendo-se que somente a modalidades preferencial foi mostradas e descrita e que todas as equivalentes, alterações e modificações conforme descritas neste documento e/ou pelas seguintes reivindicações devem ser protegidas.

[00076] Portanto, o escopo apropriado da presente invenção deve ser determinado apenas pela interpretação mais ampla das reivindicações anexas por forma a abranger todas as tais modificações, bem como todas as relações equivalentes às aquelas ilustradas nas figuras e descritas na especificação.

REIVINDICAÇÕES

1. Método implementado por computador para detectar palavras-chave predeterminadas, em um detector de palavra-chave (100) compreendendo pelo menos dados/palavras-chave de usuário (105), um modelo de palavra-chave (110), fontes de conhecimento (115) que incluem modelos probabilísticos das relações entre os eventos acústicos e pronúncias, as fontes de conhecimento compreendem um modelo acústico (120) e um preditor/dicionário de pronúncia (125), uma calculadora de recurso front-end (135), um mecanismo de reconhecimento de fala (140), e um módulo de relatório para relatar palavras-chave encontradas em tempo real (145), o método **caracterizado** pelo fato de que inclui as etapas de:

a) desenvolver o modelo de palavra-chave (110) para as palavras-chave predeterminadas pela concatenação de fonema de modelos ocultos de Markov para cada palavras-chave predeterminadas para isolar um ou mais modelos monofone, o um ou mais modelos monofone sendo selecionados de um conjunto de modelos monofones abrangendo um espaço de recurso de áudio, em que cada fonema tem três estados presentes que consiste em uma porção inicial do som (210), uma porção central do som (211) e porção final do som (212), e em que cada porção de som tem uma transição (213) dentro da própria porção ou entre porções;

b) comparar o modelo de palavra-chave (110) e o fluxo de áudio (130) para reconhecer candidatos dentre as palavras-chave predeterminadas no fluxo de áudio (130);

c) computar uma probabilidade de que uma porção do fluxo de áudio (130) corresponde a uma das palavras-chave predeterminadas do modelo de palavra-chave, em que a probabilidade é determinada utilizando uma abordagem de probabilidade com base posterior que compreende a aplicação da equação matemática:

$$P(T_i|x) = \frac{P(x|T_i)}{\sum_k P(x|M_k)}$$

onde *P* representa a probabilidade com base posterior de que um modelo T_i dentre os modelos de palavra-chave (110) corresponde a um vetor Coeficiente Cepstrais de Frequência Mel x da porção do fluxo de áudio (130), e M_k representa modelos monofone em um espaço de áudio;

d) comparar a probabilidade computada de uma palavra-chave corresponder a um limite predeterminado e declarar uma palavra detectada potencial se a probabilidade computada for maior do que o limite predeterminado;

e) computar dados adicionais para auxiliar na determinação de incompatibilidades, em que os referidos dados adicionais compreendem métricas empíricas;

f) usar os dados adicionais para determinar se a palavra detectada potencial é um alarme falso; e

g) relatar, em tempo real por um módulo de relatório, uma palavra-chave detectada se uma incompatibilidade não for identificada na etapa (f), em que o relatório compreende a geração de um relatório que é apresentado como um horário de início e término das palavras-chave detectadas no fluxo de áudio (130) com a probabilidade computada de que a palavra-chave foi encontrada.

2. Método, de acordo com a reivindicação 1, **caracterizado** pelo fato de que a etapa (b) compreende:

b.1) converter o fluxo de áudio (130) em uma sequência de características espectrais; e

b.2) comparar os modelos de palavra-chave (110) à sequência de características espectrais.

3. Método, de acordo com a reivindicação 2, **caracterizado** pelo fato de que a etapa (b.1) compreende:

b.1.1) converter o fluxo de áudio (130) em uma sequência de janelas; e

b.1.2) calcular um conjunto de 13 Coeficientes Cepstrais de Frequência Mel e seus derivados de primeira e segunda ordem para cada janela.

4. Método, de acordo com a reivindicação 1, **caracterizado** pelo fato de que a etapa (c) compreende executar um algoritmo de Viterbi.

5. Método, de acordo com a reivindicação 1, **caracterizado** pelo fato de que a etapa (c) compreende:

c.1) atribuir uma probabilidade predeterminada constante para as porções do fluxo de áudio (130) que não são compatíveis com a palavra-chave.

6. Método, de acordo com a reivindicação 1, **caracterizado** pelo fato de que o fluxo de áudio (130) compreende um fluxo de discurso falado contínuo.

7. Método, de acordo com a reivindicação 1, **caracterizado** pelo fato de que a etapa (a) compreende:

a. 1) criar um dicionário de pronúncia (125) que define uma sequência de fonemas para cada uma das palavras-chave predeterminadas;

a. 2) criar um modelo acústico (120) que estatisticamente modela uma relação entre propriedades textuais dos fonemas para cada uma das palavras-chave predeterminadas e propriedades de fala dos fonemas para cada uma das palavras-chave predeterminadas; e

a. 3) concatenar modelos acústicos para a sequência de fonemas para cada uma das palavras-chave predeterminadas.

8. Método, de acordo com a reivindicação 7, **caracterizado** pelo fato de que a etapa (a.2) compreende criar um conjunto de modelos de mistura Gaussianos.

9. Método, de acordo com a reivindicação 7, **caracterizado** pelo fato de que a etapa (a.2) compreende criar o modelo acústico (120) selecionado dentre o grupo consistindo em: modelo independente de contexto, modelo dependente do contexto e modelo trifone.

10. Método, de acordo com a reivindicação 1, **caracterizado** pelo fato de que a etapa (e) compreende computar dados adicionais selecionados dentre o grupo consistindo em: pontuações de anti-compatibilidade de palavra, porcentagem de incompatibilidade de fonema, porcentagem de compatibilidade de fonema, probabilidade de duração penalizada e um valor de certeza predeterminado.

11. Método, de acordo com a reivindicação 10, **caracterizado** pelo fato de que o valor de certeza predeterminado é escolhido para cada uma das palavras- chave predeterminadas a fim de obter uma taxa de alarme falso e precisão desejadas.

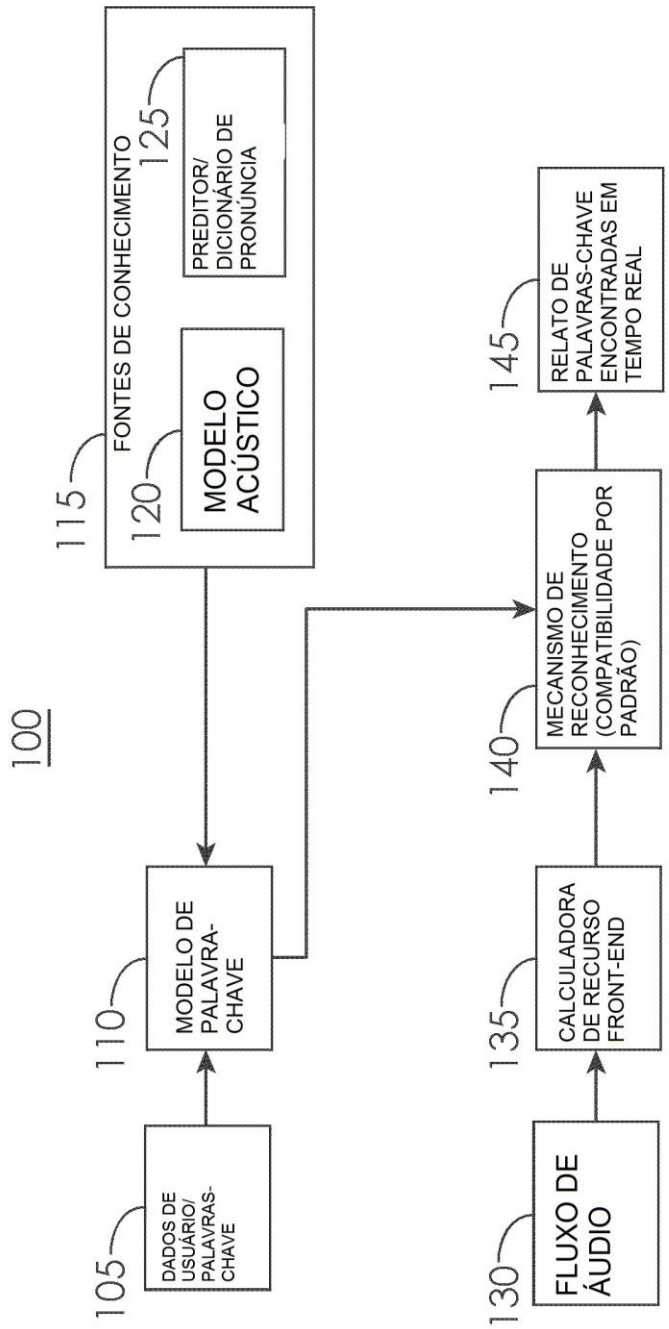
100
105
DADOS DE USUÁRIO/ PALAVRAS-CHAVE
110
MODELO DE PALAVRA-CHAVE
115
FONTES DE CONHECIMENTO
120
MODELO ACÚSTICO
125
PREDITOR/ DICIONÁRIO DE PRONÚNCIA
130
FLUXO DE ÁUDIO
135
CALCULADORA DE RECURSO FRONT-END
140
MECANISMO DE RECONHECIMENTO (COMPATIBILIDADE POR PADRÃO)
145
RELATO DE PALAVRAS-CHAVE ENCONTRADAS EM TEMPO REAL

Fig. 1

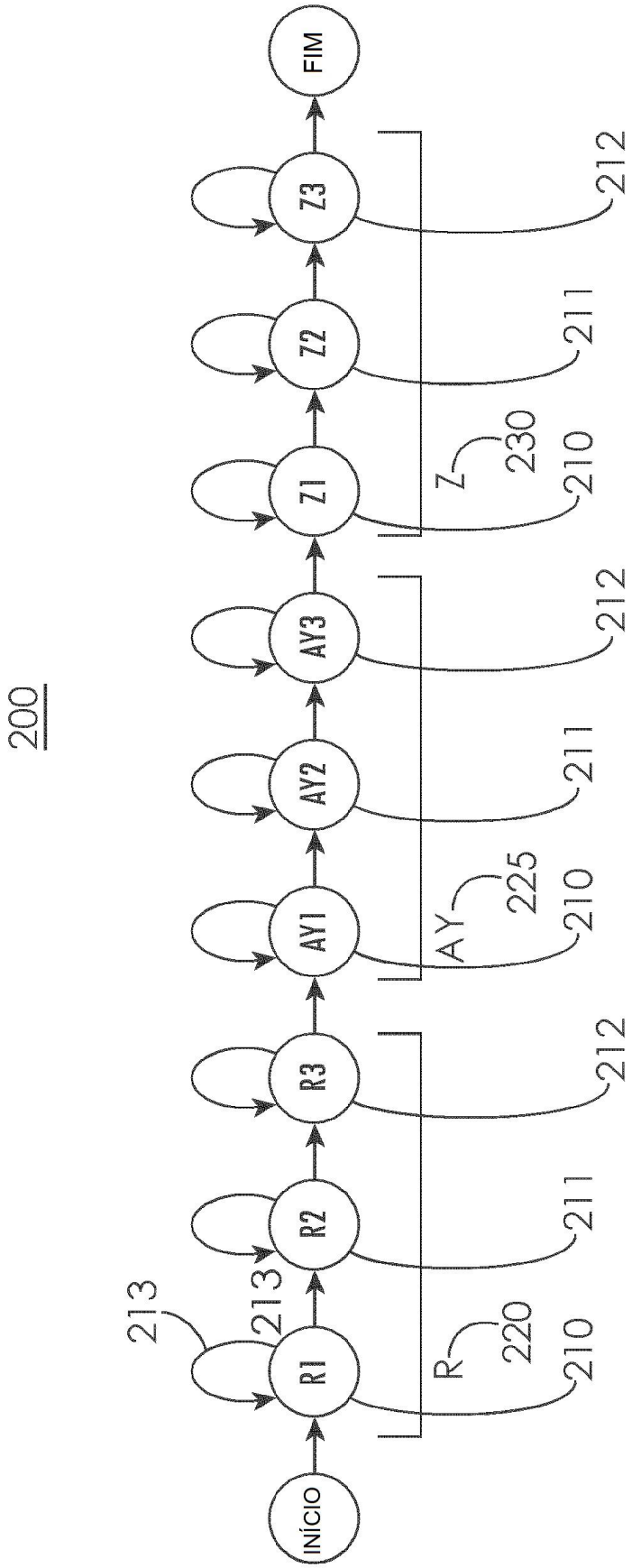
200
INÍCIO
R1
R2
R3
AY1
AY2
AY3
Z1
Z2
Z3
FIM
213
213
R
220
210
211
212
AY
225
210
211
212
Z
230
210
211
212

Fig. 2

301

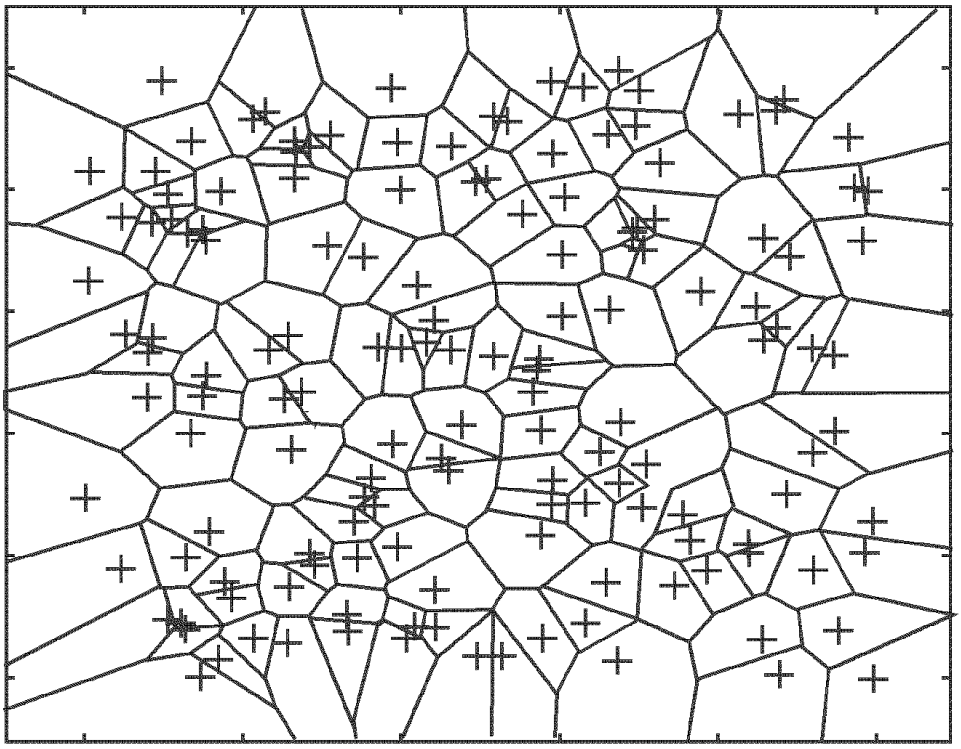

Fig. 3a

302

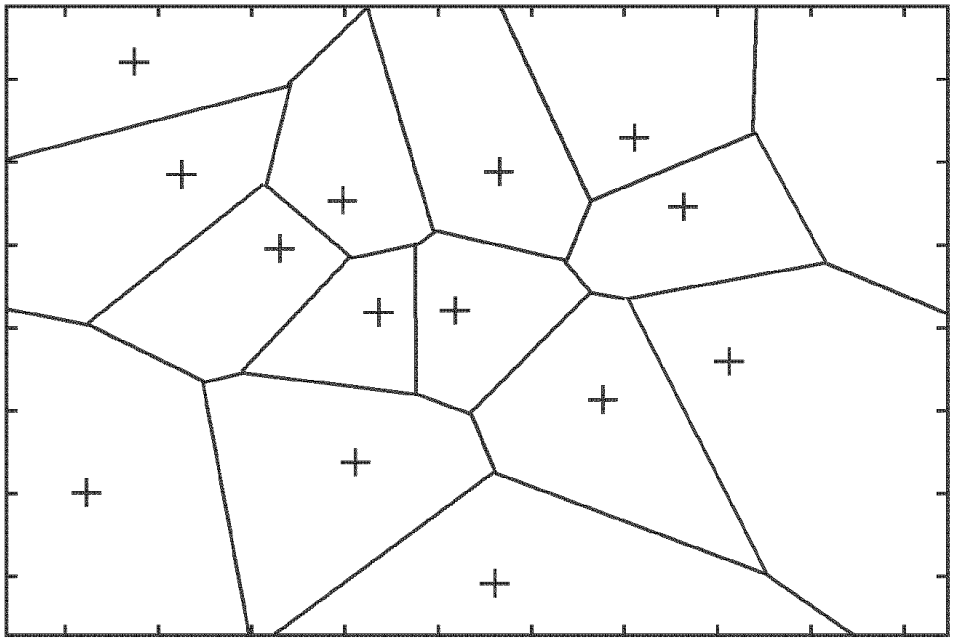

Fig. 3b

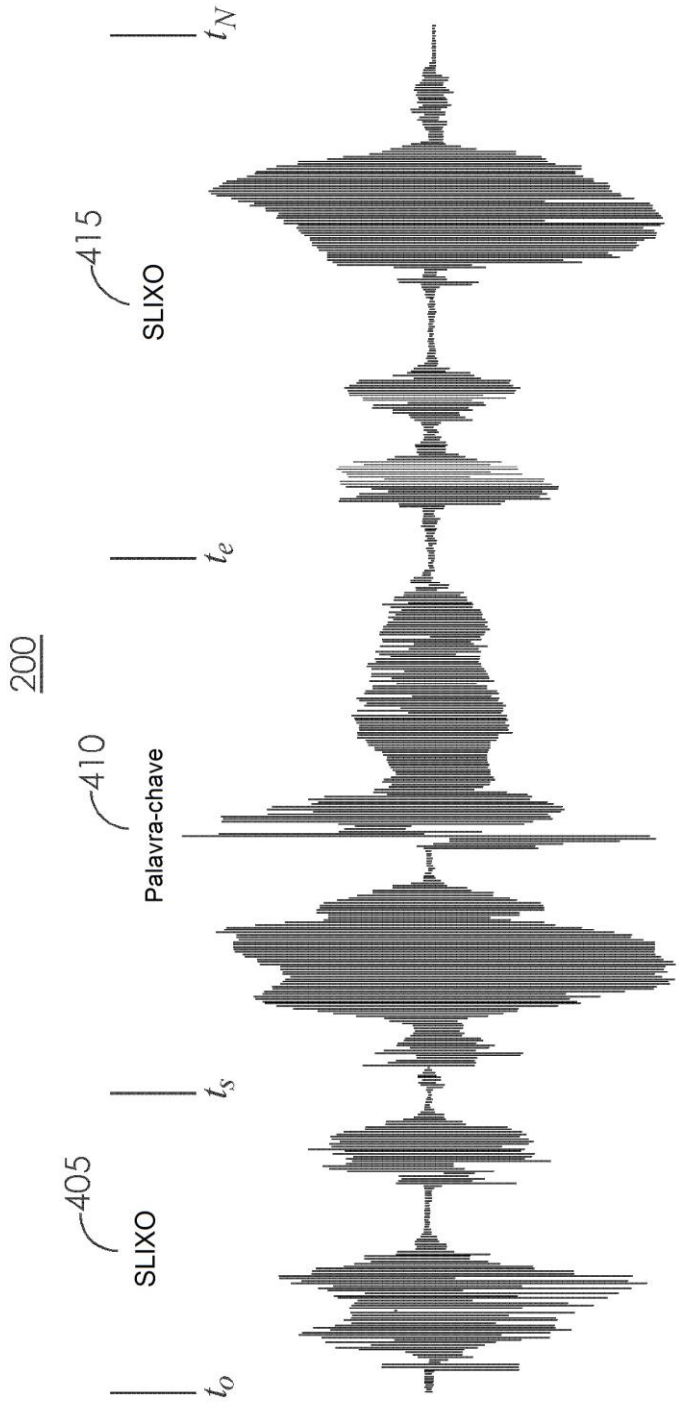

Fig. 4

500

Palavra falada	d	ih	s	eh	M	b	er
December	-0.37	-0.28	-1.17	-0.01	-0.001	-0.007	-0.002
Discover	-1.18	-1.04	-4.2	-1.5	-0.22	-0.06	-0.02

Fig. 5

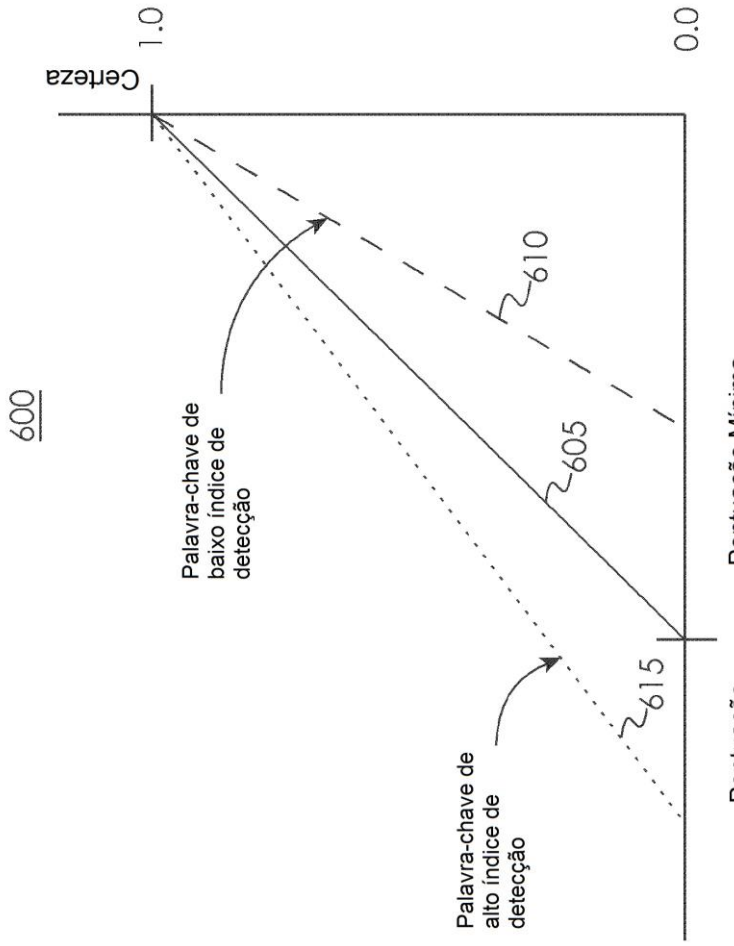
600
Certeza
1.0
0.0
Pontuação Mínima
Pontuação
610
605
615
Palavra-chave de baixo índice de detecção
Palavra-chave de alto índice de detecção

Fig. 6

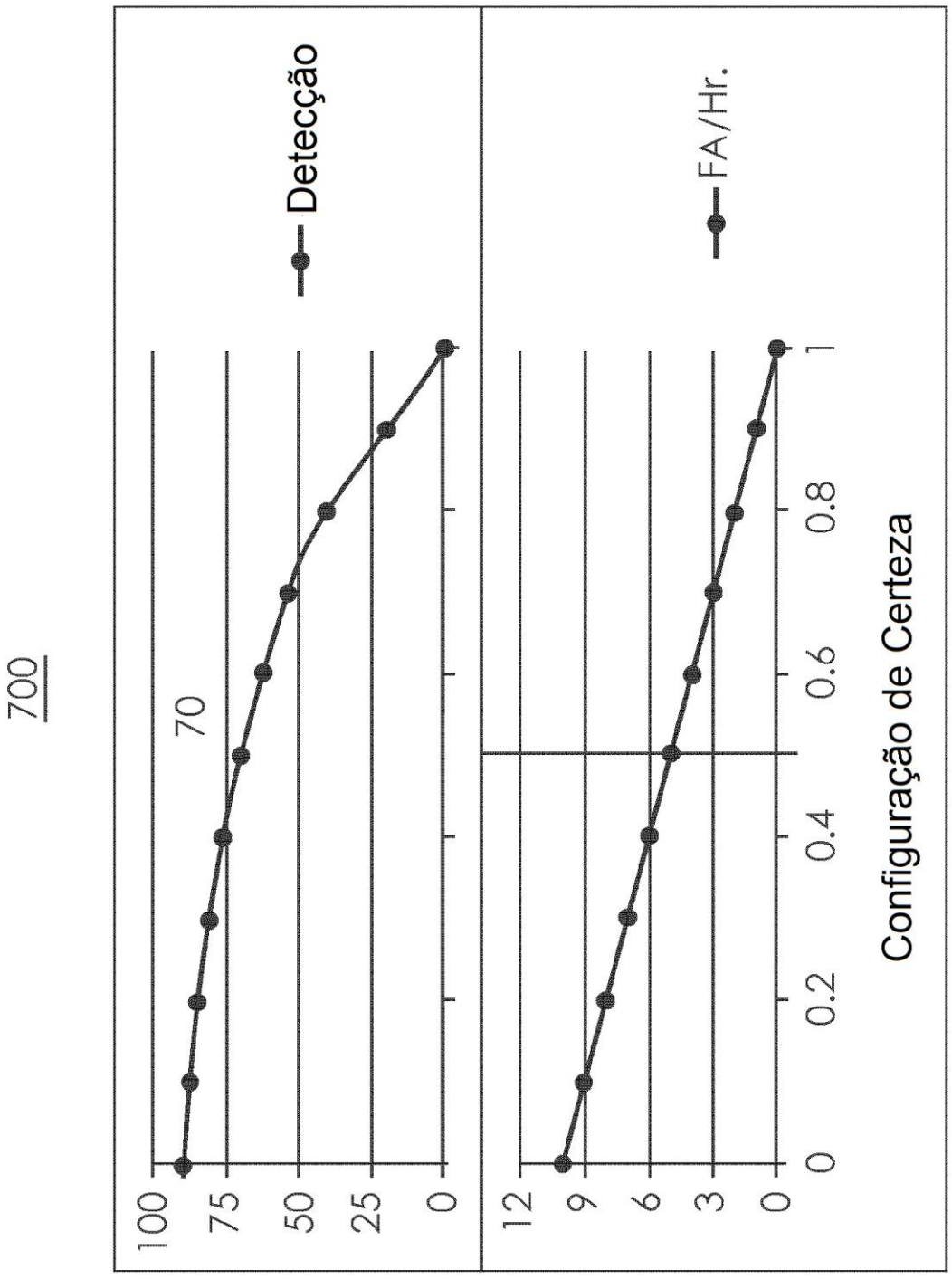
700
70
Detecção
100
75
50
25
0
FA/Hr.
12
9
6
3
0
0
0.2
0.4
0.6
0.8
1
Configuração de Certeza

Fig. 7

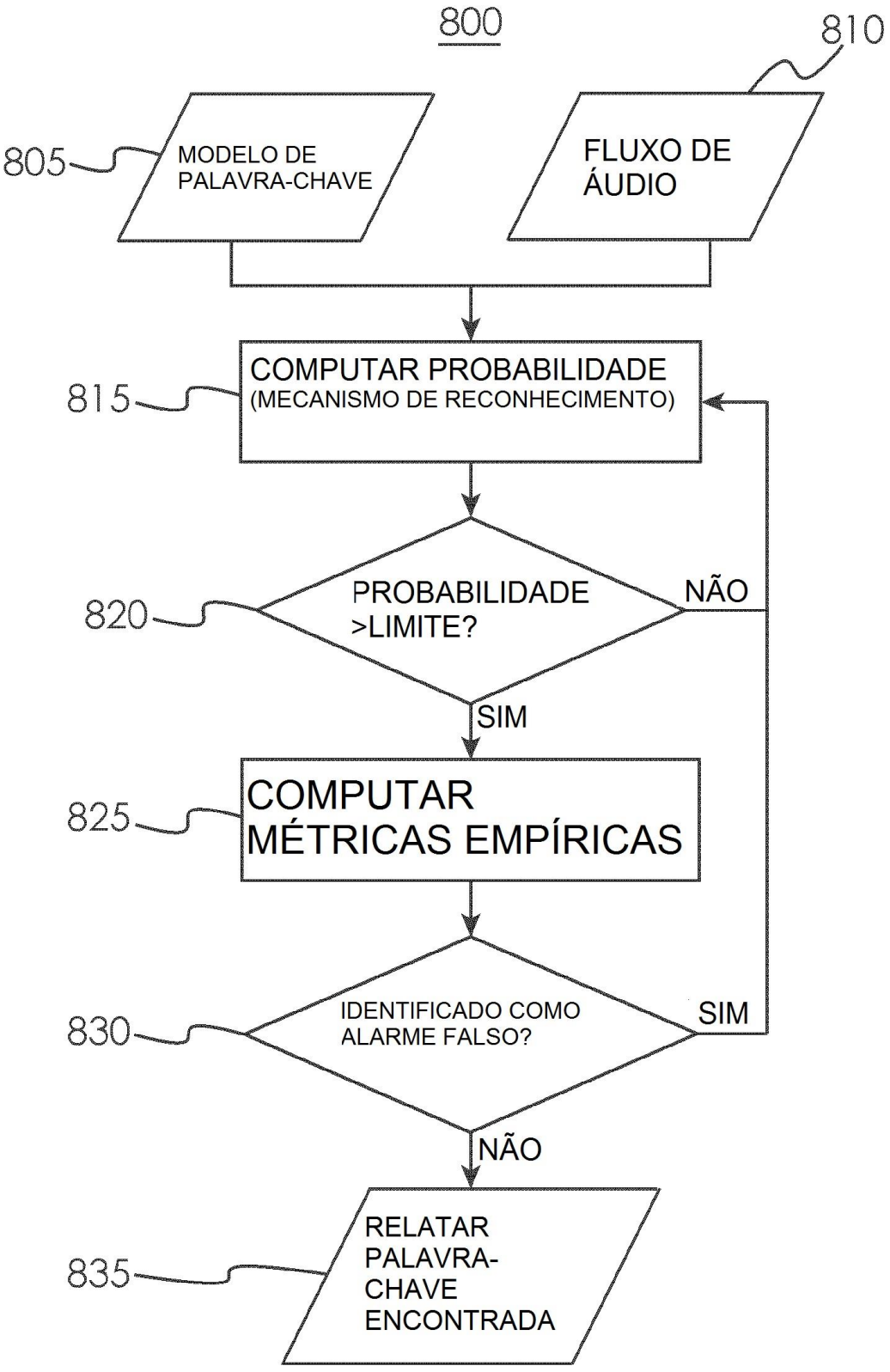
800
810
805
MODELO DE PALAVRA-CHAVE
FLUXO DE ÁUDIO
COMPUTAR PROBABILIDADE (MECANISMO DE RECONHECIMENTO)
815
PROBABILIDADE >LIMITE?
820
NÃO
SIM
COMPUTAR MÉTRICAS EMPÍRICAS
825
IDENTIFICADO COMO ALARME FALSO?
830
SIM
NÃO
RELATAR PALAVRA-CHAVE ENCONTRADA
835

Fig. 8