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

(21) **PI0904540-6 A2**

BRPI0904540A2

(22) Data de Depósito: 27/11/2009
(43) Data da Publicação: 12/07/2011
(RPI 2114)

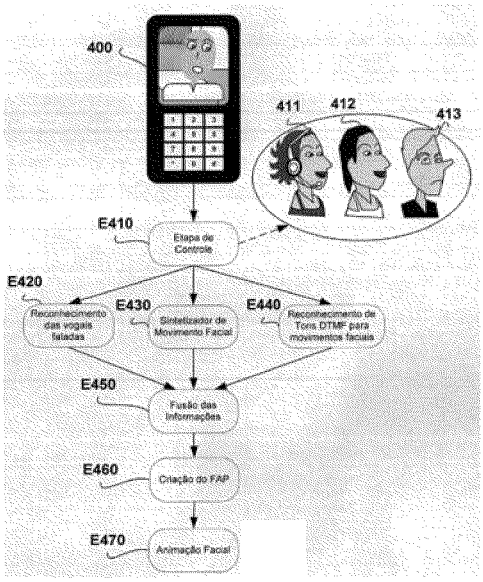
(51) Int.Cl.:
G06T 13/00 2011.01
H04W 4/18 2011.01

(54) Título: **MÉTODO DE SÍNTESE DE MOVIMENTO LABIAL PARA ANIMAÇÃO DE CABEÇAS VIRTUAIS ATRAVÉS DO PROCESSAMENTO DE VOZ EM DISPOSITIVOS PORTÁTEIS**

(73) Titular(es): Samsung Eletrônica da Amazônia Ltda

(72) Inventor(es): Antonio Carlos da Silva Barros, Cincinato Furtado Leite Neto, César Lincoln Cavalcante Mattos, Fabio Cisne Ribeiro, Francisco José Marques Anselmo, Jose Marques Soares, Paulo Cesar Cortez, Raphael Torres Santos Carvalho, Robson da Silva Siqueira, Rodrigo Carvalho Souza Costa

(57) Resumo: MÉTODO DE SÍNTESE DE MOVIMENTO LABIAL PARA ANIMAÇÃO DE CABEÇAS VIRTUAIS ATRAVÉS DO PROCESSAMENTO DE VOZ EM DISPOSITIVOS PORTÁTEIS. Aspectos da presente invenção especificam um método de comunicação utilizando um humanóide virtual animado durante chamadas convencionais. De acordo com a presente invenção, a animação é feita utilizando um sistema de reconhecimento das vogais faladas para animação dos lábios, associado ao reconhecimento de tons DTMF para animação dos movimentos da cabeça e feições faciais. Esse sistema se diferencia dos sistemas conhecidos por usar técnicas computacionalmente eficientes. Isto implica em um trabalho de otimização e ajuste de técnicas de processamento digital de sinais para viabilizar sua implementação em dispositivos portáteis. O método aqui descrito pode ser implementado de forma genérica em dispositivos portáteis como PDAs, celulares e Smart Phones que possuam acesso ao serviço de telefonia móvel.



Relatório Descritivo da Patente de Invenção para: **"MÉTODO DE SÍNTESE DE MOVIMENTO LABIAL PARA ANIMAÇÃO DE CABEÇAS VIRTUAIS ATRAVÉS DO PROCESSAMENTO DE VOZ EM DISPOSITIVOS PORTÁTEIS".**

Campo da Invenção

A presente invenção refere-se a um método de comunicação áudio-visual que utiliza dispositivos portáteis. A comunicação é feita através de um humanóide virtual 3D animado a partir de técnicas de reconhecimento de padrões de voz aplicadas a um canal de áudio.

Antecedentes da Invenção

Atualmente, muitos sistemas interativos utilizam faces falantes para se comunicar com o usuário final. Por exemplo, aplicações como secretárias eletrônicas, contadores de histórias e realidade virtual têm ganhado mais atenção através da reprodução da voz sincronizada com um movimento facial realista.

Personagens animados por computador podem ser representados em duas ou três dimensões. Conhecidos como humanóides virtuais ou *avatares*, os quais podem ser controlados por diferentes técnicas. Por exemplo, é possível animar um *avatar* através de comandos presentes em interfaces gráficas, nas quais o usuário deve escolher os

comandos dentre um conjunto finito de botões ou através de um mouse ou teclado.

A codificação MPEG4 fornece meios para implementar um humanóide virtual. Nesta codificação, existem parâmetros especializados que possibilitam a geração e transmissão do vídeo de uma "cabeça falante" sintética para fins de comunicação multimídia.

A codificação MPEG4 inclui um conjunto de Parâmetros de Animação Facial (FAP - *Facial Animation Parameters*). Estes parâmetros foram concebidos a partir do estudo de pequenas ações faciais, sendo relacionados ao movimento realizado pelos músculos da face. Esta codificação é capaz de reproduzir expressões faciais e movimentos da cabeça realizados por uma pessoa.

As referidas expressões podem ser agrupadas em duas classes: simples e complexas. Exemplos das primeiras expressões são: piscar os olhos, abrir e fechar a boca, levantar as sobrancelhas. Expressões complexas representam emoções como, por exemplo, feliz, triste, assustado.

A representação visual de um fonema constitui um visema. Os visemas são utilizados para a animação facial sincronizada com a fala, ou seja, o formato do lábio, enquanto um fonema é pronunciado.

Vários métodos de visão artificial utilizam as características de pigmentação dos lábios para realizar a sua detecção e a partir da segmentação, avaliar a forma do lábio para reconhecer o visema.

Entretanto, o contraste existente entre as cores dos lábios (não adornados) e da região facial é muito pequeno. Isto dificulta a etapa de segmentação dos lábios e torna o contorno do mesmo muito impreciso e conseqüentemente extrair as características dos lábios não se mostra eficiente. Por este motivo, o reconhecimento da forma da boca através de técnicas de visão computacional é uma tarefa complexa. Além disso, com os lábios adornados (isto é, com o uso de batom, por exemplo) se torna ainda mais complexo devido a uma grande variedade de cores existentes, dificultando, ainda mais, a concepção de um sistema automático para identificação de visemas.

Existem ainda outras dificuldades adicionais, que estão relacionadas à qualidade da imagem adquirida pela câmera digital. No caso particular das câmeras integradas a dispositivos portáteis, como é o caso dos telefones celulares, *Smart Phones* e PDAs, o tempo de exposição dos elementos sensores deixa a imagem obtida "borrada" devido ao movimento. Por esse motivo, para se conseguir uma boa definição dos movimentos da boca, faz-se necessário que a

boca ocupe uma grande porção na imagem para possibilitar uma estimação eficiente da forma dos lábios. Ao fazer isso a câmera acaba não visualizando outras partes importantes da face que são muito importantes para a comunicação.

Por isso, um sistema automático para reconhecimento de formato labial exige um custo computacional elevado para realizar as etapas de detecção e identificação das formas. Em qualquer dispositivo eletrônico, um alto custo computacional causa um aumento de consumo de energia e uma maior produção de calor.

Em dispositivos portáteis, um alto consumo de energia faz com que a bateria se descarregue mais rápido e seu uso prolongado causa uma diminuição na vida útil da bateria, visto que qualquer bateria tem o número finito de recargas. Por exemplo, uma bateria de um dispositivo portátil pode durar cerca de 300 horas em *standby* (apenas o aparelho ligado) e 7 horas em conversação.

Como o custo computacional para processar o vídeo é bem maior do que o necessário para fazer uma chamada convencional, espera-se que a duração da mesma seja bem inferior, podendo chegar a no máximo 2 horas de utilização.

Por causa dos problemas supracitados, os métodos baseados em visão artificial se concentram em apenas detectar a boca, por exemplo, aberta ou fechada. Uma vez

que a percepção da fala não depende apenas da informação acústica, o formato da boca auxilia na inteligibilidade da fala. Por exemplo, em ambientes ruidosos, o formato da boca pode compensar alguma perda de uma sílaba no canal de áudio.

Desta forma, uma maneira de tornar mais realista a comunicação através de um humanóide virtual é utilizar a voz para animar o movimento da boca, deixando os outros gestos faciais (piscar os olhos, modificação do olhar e da sobrancelha) a cargo do reconhecimento de tons DTMF.

Uma animação visual eficiente do movimento feito pela boca é útil para muitas aplicações, como, por exemplo, o treinamento da fala de pessoas com dificuldades de audição, produção de jogos e filmes, formas de interação através de agentes virtuais e comércio eletrônico.

Os métodos para o desenvolvimento deste tipo de animação são baseados em parâmetros matemáticos, nas características físicas da face, em visão artificial e no processamento de áudio.

Um exemplo de metodologia para rastreamento do movimento labial através de visão computacional foi proposto por A. W. Senior no trabalho intitulado "*Face and Feature Finding for a Face Recognition System*" publicado nos anais do "*International Conference on Audio and Video-*

based Biometric Person Authentication" p. 154 - 159 em março de 1999. Nesse trabalho, é feita uma busca da região da face utilizando um conjunto de janelas modelos de candidatos de face e de características faciais. Através de uma análise piramidal (multi-resolução), obtida através da escala das janelas modelos, é localizada a face e em seguida o processo é repetido para encontrar os elementos faciais (olhos, boca, nariz e orelhas). Uma informação extraída através deste método é um conjunto dos quatro pontos de canto da boca. Através destes, são identificadas a largura e a altura da boca que podem ser utilizadas como parâmetros para definir a sua forma, podendo ser usados para animar um humanóide virtual. No entanto, esta técnica não é vantajosa devido à quantidade de combinações de janelas realizadas para encontrar a face e os elementos faciais, tornando estes métodos complexos computacionalmente, o que dificulta a implementação em dispositivos portáteis devido ao seu limitado poder de processamento.

O documento de patente brasileiro PI 9909611-0, titular: Eyematic Interfaces, Inc, publicado em 21 de outubro de 1999 descreve um método de reconhecimento de feições para animação de um *avatar* baseado em *Wavelets*. Este documento utiliza uma série de *Wavelets* para detectar

pontos extremos da boca e, a partir destes, realiza-se o rastreamento do movimento labial. Cada ponto extremo da boca é encontrado a partir da aplicação de uma Transformada *Wavelet* com característica específica. Como de conhecimento ordinário para uma pessoa versada na técnica, para aplicar uma *Wavelet*, é necessário fazer várias convoluções durante a etapa de identificação de pontos importantes da face. Para calcular a convolução, em cada ponto da imagem, uma grande quantidade de multiplicações e somas se faz necessária. Isto torna o método muito complexo para ser utilizado em dispositivos portáteis, devido a sua limitação de memória e poder de processamento.

O artigo proposto por M-T Yang et. al. intitulado "*Lip Contour Extraction for Language Learning in VEC3D*" publicado no *Journal of Machine Vision and Applications* em abril de 2008 utiliza a segmentação dos lábios através de contornos ativos. No entanto, esse método é bastante robusto, e a procura inicial do contorno ativo e as interações subsequentes podem levar muito tempo. Em aplicações como as de vídeo-chamada, nas quais o movimento do *avatar* deve ser sincronizado com o som, esta abordagem não deve ser utilizada devido à longa duração dos procedimentos de procura e interações subsequentes.

Devido ao fato do formato labial ser o principal responsável pela formação das vogais e estas serem os principais componentes da sílaba, o reconhecimento das vogais através do processamento de voz é capaz de identificar eficientemente o formato labial e por consequência animar o humanóide virtual.

Um trabalho de reconhecimento de voz que está relacionado ao movimento facial foi proposto por D. V. McAlister *et. al* intitulado "*Lip Synchronization for Animation*", publicado nos anais do SIGGRAPH em janeiro de 1997. Este método aplica a Transformada rápida de Fourier (FFT) para extrair as características da voz e, a partir destas, realiza a animação do movimento labial. Dependendo do tempo de aquisição e da taxa de amostragem do sinal, este método pode se tornar custoso computacionalmente, e por isso, não é vantajoso para aplicação em dispositivos portáteis com baixo poder computacional, como os dispositivos em utilização preferencial na presente invenção.

Um método similar foi proposto por G. Zoric e I. S. Pandzic no trabalho intitulado "*Real-time Language Independent Lip Synchronization Method Using a Genetic Algorithm*" publicado no *Journal of Signal Processing*, p. 3644 a 3656 em dezembro 2006. Nesse trabalho, o resultado

da Transformada rápida de Fourier (FFT) é convertido em uma nova escala. No sinal convertido, é aplicada a Transformada discreta do cosseno (DCT) e, após todas estas etapas, são extraídos os coeficientes que representam o movimento labial. Para aplicações com processadores dedicados ou em ambiente PC, o método é capaz de operar em tempo real. Contudo, a quantidade de operações necessárias para realizar este procedimento é muito maior do que o método proposto por McAlister, tornando-se inviável para aplicações em dispositivos portáteis devido ao custo computacional de todas estas operações.

Na patente US 6,735,566, concedida em 11 de maio de 2004, é proposto um método que utiliza o reconhecimento de voz para uma animação facial realista. Este método utiliza um treinamento associando o vídeo da boca à voz para modelar o movimento labial. Este método utiliza um Modelo Escondido de Markov (*Hidden Markov Model*) para a extração das características labiais de cada som falado. Este tipo de abordagem possui altas taxas de acerto e uma grande confiabilidade, no entanto, é um método de reconhecimento de padrões muito complexo computacionalmente, o que o torna impraticável devido ao custo computacional elevado.

Outro exemplo de animação facial a partir da voz foi descrito na patente US 6,665,643, concedida em 16 de

dezembro de 2003, titular: Telecom Italia Lab S.P.A. De acordo com os ensinamentos da presente invenção, o reconhecimento de fonemas (vogais e consoantes) falados é realizado para animar um modelo virtual. Na referida patente, cada palavra falada é transformada em um texto e a partir do texto, são identificados os fonemas. A referida solução se mostra bastante eficiente, contudo requer o reconhecimento de muitos fonemas. O melhor desempenho é obtido identificando o conteúdo de toda a mensagem falada, sendo indicado para comunicação *off-line*.

O artigo proposto por S. Kshiragas e N. Magnenat-Thalmann intitulado "*Lip Synchronization Using Linear Predictive Analysis*" publicado no IEEE em julho de 2000 realiza o reconhecimento de vogais utilizando a codificação por predição linear (LPC - *Linear Predictive Coding*') para a extração de características e estas são processadas por uma rede neural.

Um método similar foi proposto por O. Farooq e S. Datta em seu trabalho intitulado "*Phoneme Recognition using Wavelet Based Features*" publicado no *Journal of Information Sciences* vol. 150, p. 5 - 15 em março de 2003. Este utiliza a Transformada rápida *Wavelet* para extrair as características do sinal de áudio e também utiliza uma rede neural para reconhecer fonemas na língua inglesa.

A extração de características por predição linear ou *Wavelet* seguida de sua aplicação em uma rede neural possui baixa complexidade computacional. Em ambos os métodos, o reconhecimento das vogais é feito para falantes da língua inglesa. Contudo, é importante ressaltar que a pronuncia em outras línguas, como, por exemplo, a língua portuguesa, possui uma variedade de fonemas muito maior. Isso se deve ao fato de que uma mesma vogal pode ter algumas variações tônicas e nasais graças aos diferentes sotaques das diversas regiões brasileiras. Consequentemente, os métodos baseados em predição linear ou *Wavelet* possuem o inconveniente de gerar falsos reconhecimentos devido a esta variedade.

O documento de patente US *20090158184*, titular: AOL LLC, publicado em 18 de junho de 2009 reivindica um sistema e um método para animar um *avatar* com base na animação percebida em um segundo *avatar*, o método compreendendo as etapas de representar graficamente um primeiro usuário com um primeiro *avatar* capaz de ser animado; representar graficamente um segundo usuário com um segundo *avatar* capaz de ser animado, em que mensagens de comunicação são enviadas entre o primeiro e o segundo usuário; receber uma indicação de uma animação de um primeiro *avatar*; acessar informação associando animações com o *avatar*; identificar,

com base na informação acessada, uma animação para o segundo *avatar* que é responsiva à animação indicada do primeiro *avatar* e em resposta à indicação recebida, animar o segundo *avatar* com base na animação responsiva identificada. De acordo com os ensinamentos do referido documento de patente, o *avatar* é animado através de uma aplicação tipo mensagens on-line (como, por exemplo, MSN ou Skype). O *avatar* se mexe em função das palavras escritas no sistema. Assim, não há qualquer tipo de reconhecimento de sons.

A patente US 7,176,956, concedida em 13 de fevereiro de 2007, titular: MOTOROLA INC, trata da animação de *avatares* em uma comunicação entre dispositivos portáteis (vídeo chamada). Os *avatares* se mexem através das mudanças de parâmetros obtidos através de técnicas de reconhecimento de imagens providas pela câmera do celular.

A patente US 7,231,205, concedida em 12 de junho de 2007, titular: Telefonaktiebolaget LM Ericsson trata da animação de *avatares* em uma comunicação entre dispositivos portáteis. O sistema é conectado a um servidor que promove o enlace entre os dispositivos e é ele que provê o serviço de *avatares*. O estado dos *avatares* pode ser modificado através do teclado do celular, mas não prevê o reconhecimento de sons.

A patente US 6,665,640, concedida em 16 de dezembro de 2003, titular: Phoenix Solutions, Inc apresenta um *avatar* animado utilizando a voz. O *avatar* usa FAPs como parâmetros de movimentação. Os FAPs são obtidos diretamente de um stream MPEG4. Este sistema não faz a simplificação dos visemas, nem é otimizado para dispositivos com pouco poder de processamento, como os telefones móveis atuais.

A patente US 7123262, concedida em 17 de outubro de 2006, titular: Telecom Italia Lab S.p.A usa visemas e gera FAPs sobre um rosto previamente parametrizado com *Active Shape Model*. De acordo com o referido documento, a voz e a imagem são unidas para fazer a movimentação do modelo de rosto, o que não constitui um *avatar*, mas sim uma técnica de animação de um rosto modelado. Estas técnicas são, em geral, pesadas e complexas, o que inviabiliza a aplicação das mesmas em dispositivos portáteis.

O documento WO 2008031955, publicado em 20 de março de 2008, descreve um método e sistema para a animação de um *avatar* em um aparelho móvel com base no sinal de som correspondendo à voz de um interlocutor em uma conversação telefônica. O referido método propõe a aparência e movimentação dos avatares em tempo real ou quase real, sendo o avatar escolhido e/ ou configurado através de um serviço online na rede. O sistema do documento WO

2008031955 compreende um aparelho de comunicação móvel, servidor de recepção de sinal, bem como meios de cálculo e análise do sinal de som para movimentar o avatar e simular conversação em tempo real.

Aspectos da presente invenção especificam um método de comunicação utilizando um humanóide virtual animado durante chamadas convencionais. De acordo com a presente invenção, a animação é feita utilizando um sistema de reconhecimento das vogais faladas para animação dos lábios, associado ao reconhecimento de tons DTMF para animação dos movimentos da cabeça e feições faciais. Esse sistema se diferencia dos sistemas conhecidos por usar técnicas computacionalmente eficientes. Isto implica em um trabalho de otimização e ajuste de técnicas de processamento digital de sinais para viabilizar sua implementação em dispositivos portáteis.

O método aqui descrito pode ser implementado de forma genérica em dispositivos portáteis como PDAs, celulares e *Smart Phones* que possuam acesso ao serviço de telefonia móvel.

A presente invenção se diferencia das soluções conhecidas por propiciar as seguintes vantagens diante do estado da técnica:

- Baixo custo computacional: os processamentos realizados possuem um baixo esforço computacional e podem

ser utilizados para comunicação em tempo real, portanto, em dispositivos portáteis.

- Independência à intensidade da fala: a animação da abertura da boca é feita a partir da avaliação da energia dos últimos segundos da comunicação. Independentemente da intensidade (alta ou baixa) do sinal de áudio, o *avatar* move os lábios naturalmente.

- Reconhecimento de vogais preferencialmente adaptado para o reconhecimento do idioma português: o método proposto foi adaptado para reconhecer as vogais sob condições diferentes de sotaques regionais, como ocorre, por exemplo na língua portuguesa.

- Gerador sintético de expressões faciais: o método proposto possui a capacidade de gerar expressões faciais sintéticas de forma próxima às expressões faciais reais.

- Movimentação do humanóide virtual através de tons DTMF: o sistema proposto possibilita a reprodução de movimentos da cabeça e expressões faciais utilizando o canal de voz, não requisitando o envio de informação nos canais de dados durante a comunicação.

- Novas formas de vídeo-chamada: o método proposto pela presente invenção pode ser utilizado em diferentes aplicações a fim de agregar valor ao serviço de chamadas convencionais das operadoras de telefonia móvel através de

sua utilização ao receber uma chamada originada de um telefone fixo ou móvel.

- Recurso Stand Alone: apenas um dos terminais envolvidos na conversação precisa ter um dispositivo para exibição do *avatar*. O segundo terminal pode interagir com o primeiro que contém os meios de exibição pressionando uma das teclas de um telefone convencional, por exemplo.

- Diminuição no tráfego da rede: a transmissão pelo canal de voz utiliza bem menos tráfego que uma comunicação utilizando vídeo-chamadas e deixa-o disponível para outras aplicações como, por exemplo, o vídeo *streaming*.

- Utilização de dispositivos portáteis de baixo custo: a presente invenção pode ser utilizada em celulares de baixo custo que não possuem componentes adicionais, como por exemplo, câmera, touch-scren, acesso à rede 3G.

Sumário da Invenção

O método proposto pela presente invenção realiza a reprodução do movimento labial do humanóide virtual através do reconhecimento das vogais faladas. A utilização de vogais é vantajosa, pois estas são produzidas a partir da acústica da cavidade oral, influenciando o formato do lábio.

Além disto, o sistema proposto propiciará a animação dos movimentos da cabeça e expressões faciais através da

associação desses gestos com os tons de duas freqüências utilizados na discagem dos telefones (*Dual-Tone Multi-Frequential* - DTMF). Por exemplo, ao receber o tom DTMF da tecla 2, o humanóide pisca o olho esquerdo. Os tons DTMF são formados a partir da combinação de dois impulsos elétricos de freqüências especificadas. A probabilidade de a voz humana gerar a combinação de duas freqüências utilizadas neste padrão é baixíssima, por isto, elas são simples de serem detectadas e filtradas.

Com base no princípio acima exemplificado, o sistema da presente invenção tornará mais agradável a comunicação em chamadas de áudio, visto que o sistema é capaz de passar a sensação, para o usuário, de uma conversa utilizando vídeo-chamada. A partir da utilização deste sistema, não seria necessário adquirir o vídeo, comprimi-lo e enviar os dados pela rede.

Além disto, o humanóide possibilitará às pessoas idosas e pessoas com dificuldade de audição compreender melhor a conversa, pois a percepção da fala não depende apenas da informação acústica. As informações visuais como, por exemplo, os movimentos labiais e expressões faciais influenciam na percepção da informação falada.

Um grande diferencial da invenção é que esta possibilita uma inserção de novas funcionalidades a

dispositivos portáteis e móveis, sem a necessidade de modificação no seu desenho, seja na sua placa de montagem ou pela adição de teclas extras no seu layout. Tudo isso sem a inclusão de um servidor intermediário para tratamento de dados, o que possibilita a utilização da presente invenção em dispositivos atualmente disponíveis no mercado a partir de uma simples atualização de software.

A presente invenção fornece uma nova funcionalidade aos dispositivos portáteis existentes no mercado através da sua utilização ao acessar caixa-postal, atendentes virtuais ou receber chamadas. Por exemplo, uma atendente virtual poderia interagir com o usuário dependendo dos comandos enviados pelo usuário. Neste caso, ao apertar uma tecla que não está nas opções disponíveis no serviço, o *avatar* começa a balançar a cabeça indicando ao usuário que ele fez algo errado ou fica com as feições tristes.

Para diminuir o esforço computacional e tornar viável a aplicação em um dispositivo portátil, na presente invenção, as vogais são utilizadas para animar o humanóide virtual. A vogal constitui a base da sílaba e sua emissão é feita basicamente pelo movimento labial. Por exemplo, ao fazer os lábios ficarem com a forma utilizada para pronunciar a vogal "o", não é possível pronunciar a nenhuma das outras vogais.

Além disso, a presente invenção é especialmente adaptada a um método de extração de características otimizado para idiomas com maior número de variações fonológicas, como, por exemplo, o português do Brasil, realizando adaptações para tornar robusto o método quanto a estas variações de entonações.

Breve Descrição das Figuras

Os objetivos e as vantagens da presente invenção tornar-se-ão mais evidentes a partir da descrição detalhada a seguir de um exemplo de concretização da invenção e desenhos anexos fornecidos a título de exemplo não-limitativo, em que:

A figura 1 apresenta um diagrama esquemático do funcionamento do sistema proposto.

A figura 2 apresenta um modelo genérico de dispositivo portátil, no qual a presente invenção pode ser implementada.

A figura 3 apresenta um diagrama de blocos do método para animação do humanóide virtual.

A figura 4 apresenta uma visão geral do funcionamento do método proposto pela presente invenção.

A figura 5 apresenta um diagrama detalhado da etapa de reconhecimento do formato labial e amplitude de

abertura da boca, de acordo com a modalidade preferida da presente invenção.

A figura 6 apresenta um diagrama detalhado da etapa de reconhecimento do formato labial utilizando *Wavelets* e Rede Neural, de acordo com a modalidade preferida da presente invenção

A figura 7 apresenta uma vista frontal do humanóide virtual e principais pontos característicos utilizados nesta proposta de patente de invenção para animação facial.

Descrição das Concretizações Preferidas da Invenção

Na figura 1, é ilustrado o funcionamento do sistema da presente invenção. Um usuário 100 realiza uma conversação através de chamada convencional utilizando um dispositivo portátil 101. Esta chamada pode ser efetuada entre o dispositivo portátil e um usuário 110 utilizando um telefone fixo 111 ou um telefone celular 112. Além disto, o usuário 100 pode utilizar o sistema proposto para acessar serviços da operadora de telefonia móvel 120, como, por exemplo, a caixa postal 121, o serviço de auto-atendimento 122 ou reproduzir mensagens de voz 123.

Durante a chamada, o usuário 100 tem a liberdade de iniciar o sistema proposto e passar a visualizar o humanóide virtual. A cada sílaba reconhecida, a vogal e a boca do humanóide é animada na tela 131 do dispositivo

portátil 101. Eventualmente, ao receber um tom DTMF, o dispositivo portátil 101 realiza a modificação da feição do humanóide 132 ou ativa um gesto específico 133. Caso os dois dispositivos utilizados na comunicação possuam o sistema proposto, ambos poderão ter a sensação de uma vídeo-chamada, em que cada cliente pode controlar as feições do humanóide virtual e tem a boca animada pelo sistema proposto. Isto tornará mais interessante e divertida a chamada tradicional, pois emoções e sentimentos poderão ser mais perceptíveis através da utilização do humanóide virtual, bem como auxiliará na inteligibilidade da comunicação para pessoas que possuem alguma dificuldade de audição causada pelo ruído do ambiente.

De acordo com a modalidade preferida da invenção, o dispositivo computacional deve ser composto por unidade de processamento central ou outro elemento de processamento para executar instruções computacionais com memória para armazenamento de informações e instruções, display ou outro dispositivo que exiba ou forneça saída visual de informações, teclado ou outro dispositivo de entrada para inserção de informações, componentes de entrada e saída de áudio tais como microfone e alto falante; e componentes que forneçam o acesso a rede de telefonia móvel, conforme mostrado na figura 2.

O método aqui proposto permite a animação dos movimentos da cabeça, a seleção de feições e o reconhecimento do movimento labial.

O método da presente invenção utiliza os parâmetros FAP para animar o humanóide virtual. Esses parâmetros FAP são compostos por um conjunto de 68 parâmetros que definem a modificação da forma ou movimentos faciais. O método de reconhecimento de voz da presente invenção combina uma série de algoritmos com o objetivo de aperfeiçoar seu esforço computacional e robustez, visando viabilizar a sua utilização em dispositivos com restrições de capacidade computacional, mais notadamente, os dispositivos portáteis. Este método está dividido conforme apresentado na

A figura 3 e é composto das seguintes etapas:

1. Configuração da comunicação: nesta etapa são avaliadas as opções pessoais do usuário. Este tem a liberdade de associar um *avatar* dentre um conjunto de humanóides virtuais disponíveis no dispositivo portátil que mais se assemelhe a um contato de sua agenda telefônica. Desta forma, ao receber uma chamada ou uma mensagem da caixa postal deste usuário, é perguntado ao usuário se este deseja ativar a animação do humanóide virtual. Caso o usuário deseje este tipo de comunicação, o *avatar* associado é utilizado para a comunicação. Além disto, dentre os

humanóides virtuais disponíveis, existem aqueles exclusivos para as operadoras de telefonia móvel, no qual cada operadora possui um humanóide específico e ao utilizar o sistema proposto para acessar algum serviço de auto-atendimento acessado pelo usuário de telefonia móvel.

2. Aquisição do áudio: nesta etapa, é realizada a aquisição do áudio recebido (MMS ou chamada convencional) em formato padronizado para ser utilizado pelas outras etapas de reconhecimento.

3. Análise de Energia do Áudio: o método proposto analisa a relação da energia do sinal em uma quantidade fixa de amostras do sinal de áudio. Esta quantidade forma o que chamamos de quadro de áudio. É calculada uma relação entre a energia do *quadro* atual e de seus anteriores, dentro de uma janela de tempo para dimensionar o quanto a boca está aberta.

4. Reconhecimento das vogais faladas: nesta etapa são analisados os quadros de áudio para reconhecer as vogais faladas. Esta etapa realiza a extração de características do sinal de áudio e as aplica em uma rede neural progressiva (*feed foward propagation*) com pesos fixos, gerados a partir de um treinamento feito fora do dispositivo portátil utilizando um conjunto de amostras de áudio. Este tipo de solução foi escolhido devido seu baixo

custo computacional. Um diferencial da presente invenção quanto aos métodos conhecidos é a otimização do método de extração de características para o reconhecimento das vogais faladas em diferentes entonações e sotaques do idioma português brasileiro.

5. Sintetizador de expressões faciais: nesta etapa, são gerados artificialmente alguns gestos específicos do rosto que simulam movimentos naturais feitos por uma pessoa. Por exemplo, como piscamos inúmeras vezes ao longo do dia, este movimento pode ser simulado através de um processo aleatório e usado para animar os olhos e olhar do humanóide virtual.

6. Animação da cabeça e gestos faciais: nesta etapa são reconhecidos os tons DTMF recebidos no canal de áudio. Após o reconhecimento do tom recebido, é configurada a feição ou movimentos faciais pré-definidos.

7. Fusão das Informações: nesta etapa, as informações reconhecidas e geradas artificialmente são organizadas para formar uma única informação que será utilizada durante a animação do humanóide virtual. Esta etapa realiza uma análise de prioridades entre cada informação recebida. Nas informações, a vogal reconhecida tem prioridade maior do que as feições faciais reconhecidas pelos tons DTMF. Enquanto o usuário está falando, o formato

labial é controlado pelo reconhecimento de vogais e, caso contrário, o formato é controlado a partir da feição escolhida através do tom DTMF.

8. Geração dos Parâmetros de Animação Facial: nesta etapa do processamento, as instruções são convertidas em parâmetros FAP, utilizando as informações definidas na etapa anterior. Por exemplo, a forma dos lábios é dependente de quatro pontos característicos da boca. Ao receber a instrução de boca vogal "A", um pacote FAP é gerado, no qual os quatros pontos que definem a boca são identificados e estes são enviados para a animação do humanóide.

9. Animação do humanóide virtual: nesta etapa, é realizada a modificação das feições do humanóide através dos FAP recebidos.

O método proposto pela presente invenção é apresentado na figura 4. O sistema no qual o referido método pode ser implementado é composto por um dispositivo portátil, representado pela referência 400, integrado a um método de processamento de áudio e de geração da animação do humanóide virtual.

A primeira etapa do método, representada por E410, realiza o controle da aplicação de reconhecimento de voz e pose da cabeça. Dependendo da personalização do usuário,

um *avatar* específico é utilizado para a comunicação dentro de um conjunto de *avatares* 411 a 413. Por exemplo, ao realizar uma chamada para o serviço de auto-atendimento, para pessoa do sexo feminino e para uma pessoa do sexo masculino, são usados os *avatares* 411, 412, 413, respectivamente.

Após isto, são realizadas as etapas de reconhecimento das vogais faladas E420, sintetização de movimento facial P430 e reconhecimento de tons DTMF E440. Por exemplo, quando o usuário remoto está falando, a etapa E420 realiza o reconhecimento das vogais faladas e, na etapa E430, os movimentos do olhar são animados durante toda a chamada. Ao receber um tom DTMF, na etapa E440, é realizada a modificação do tipo de feição, entre um conjunto de, por exemplo, doze feições padronizadas.

A etapa E420 é dividida em várias etapas conforme ilustrado detalhadamente na Figura 5. A primeira etapa deste processamento consiste na geração do quadro de áudio, representada por E500. Esta pode ser feita através da amostragem da voz recebida pela linha telefônica 501 ou do arquivo disponível em uma mensagem MMS 502. Em todos os casos, um quadro de áudio com duração de 32 ms é adquirido e para esta duração são processadas 512 amostras, que formam o quadro de áudio 503. Por exemplo,

independentemente da origem, o sinal de áudio é convertido para o formato PCM com 512 amostras para cada 32 ms, representado esquematicamente através de 503. Esta conversão é feita através do código implementado no DSP do dispositivo portátil.

Reconhecimento da Abertura da Boca

Após esta etapa de condicionamento de dados, é realizado o cálculo da energia do sinal de voz em cada quadro adquirido em E510. Cada valor é colocado em um *buffer* e é calculada a energia máxima nos últimos N ms representada por $\overline{E_{Max}}$, conforme ilustrado em 511, em que N varia entre 0,5 e 1,5 segundos, sendo preferencialmente utilizado o valor de 1 segundo.

Normalmente, a energia da boca varia ao pronunciar vários fonemas. Por exemplo, ao sustentar um fonema, a energia é máxima e durante as pausas entre palavras a energia é praticamente nula. Por causa disto, é calculado o mínimo de energia para a detecção de voz $\overline{E_{Min}}$ como uma fração de $\overline{E_{Max}}$. Esta razão pode variar de entre 1 % a 50 %, sendo preferencialmente utilizado o valor de 10%. Quando a energia é menor que este mínimo, indica que o usuário não está falando, mantendo o *avatar* com a boca fechada.

Caso a animação fosse feita só em função do valor RMS da energia, seria necessário definir um valor mínimo e máximo. Desta forma, uma pessoa falando baixo, faria um movimento pequeno nos lábios do *avatar*.

Esta razão entre $\overline{E_{Min}}$ e o $\overline{E_{Max}}$ possibilita um reconhecimento de abertura da boca independente da intensidade da voz. O formato da boca, independentemente do usuário falar baixo ou alto, se adapta à animação facial em função da razão $\overline{E_{Min}}$ / $\overline{E_{Max}}$.

Quando a energia é maior do que um limite especificado, o método da presente invenção realiza o processamento do quadro de áudio para identificar a amplitude da abertura da boca na etapa E520 e a vogal falada na etapa E530. Desta forma, a razão entre a energia máxima e a energia do quadro atual é utilizada para controlar a dimensão de abertura da boca. Por exemplo, quando a energia é igual a $\overline{E_{Max}}$, atribui-se o máximo de abertura, conforme mostrado na etapa 521. Quando a energia é menor que $\overline{E_{Min}}$, atribui-se a boca fechada. No momento em que a energia é maior do que $\overline{E_{Min}}$ e menor do que $\overline{E_{Max}}$, a boca é desenhada em função da razão mencionada anteriormente.

Reconhecimento das vogais faladas

Após este cálculo, na etapa E520, o formato e a dimensão de abertura da boca são determinados. O mesmo quadro de áudio é processado a fim de extrair características capazes de identificar o tipo de vogal falada e tem seu funcionamento detalhado na figura 6.

Como um sinal de voz pode ser considerado estacionário dentro de uma janela de tempo em torno de 10ms. Na etapa E600, o quadro E601 é divido em quadro sub-quadros, conforme mostrado etapa E602. Em cada um destes, são extraídas as características da voz através dos blocos E610 e E620.

A etapa E620 realiza a extração de características, utilizando, preferencialmente, uma abordagem multi-escala *Wavelet* não-padrão. De acordo com a concretização preferida da invenção, a análise de um sinal através de Transformada *Wavelet* é feita através da múltipla convolução de duas funções (escalamento e *Wavelet*), conforme a abordagem da Transformada *Wavelet* Rápida (FWT - *Fast Wavelet Transform*).

É importante mencionar que a aplicação seguida destas funções é complexa computacionalmente em uma imagem. Isto ocorre devido à grande quantidade de pixels da imagem. Já o sinal de áudio processado possui apenas 64 amostras. Mesmo

aplicando convoluções sucessivas (para o cálculo de vários níveis) o custo computacional é baixo.

A Transformada *Wavelet* Rápida realiza a decomposição de um sinal através da convolução do sinal seguido de uma sub-amostragem. Na etapa E610, é realizado o cálculo dos coeficientes em vários níveis de escalamento, conforme mostrado na etapa 611.

A energia E^j de um nível j para todos os níveis da Transformada *Wavelet* pode ser calculada a partir da soma dos quadrados dos coeficientes de detalhes.

De acordo com a concretização da presente invenção, na etapa E620, além da energia E^j, são calculadas a energia total E^{tot} e a entropia residual *Wavelet* H^j, respectivamente descritas por

$$E^{tot} = \sum_{j}^{J_{max}} E^j \quad , \tag{1}$$

$$H^j = -\frac{E^j}{E^{tot}} \cdot \log \frac{E^j}{E^{tot}} \, . \tag{2}$$

Para diminuir o esforço computacional no cálculo de extração de características e melhorar o esforço computacional, de acordo com a presente invenção, a energia é calculada para níveis específicos da decomposição. As características podem ser calculada em qualquer combinação

entre os níveis da transformada, de 1 a J_{max}, preferencialmente são utilizados os níveis 1, 3, 4 e 5 para calcular os coeficientes, sendo a escolha de não utilizar um nível específico feita através de testes experimentais com várias combinações de características para verificar a combinação que apresenta melhor desempenho.

Após este cálculo, na etapa E630, as características são aplicadas em uma rede neural progressiva, treinada com o algoritmo de *backpropagation*. Esta rede neural possui, preferencialmente, N neurônios na camada de entrada (N = número de características usadas), oito na camada escondida e seis na camada de saída, em que as cinco primeiras saídas indicam cada tipo de vogal e a sexta saída indica a ausência de vogal. Na etapa de reconhecimento, o valor da saída que possui maior intensidade é considerada como vogal falada.

Sintetizador de expressões faciais

A seguir é descrito o método de geração artificial dos movimentos dos olhos e do olhar, representado por E430. Para piscar os olhos, o perfil de olho fechado é aplicado na etapa E430, espera-se 100 milissegundos e aplica-se o perfil de olho aberto novamente simulando o piscar dos olhos. Em seguida, seleciona-se um valor inteiro aleatório

entre 3000 e 4500 e o usa-se como tempo, em milissegundos, entre a piscada atual e a próxima, sendo o procedimento repetido.

Para controlar os movimentos do globo ocular é feito um processamento semelhante. Neste caso, o controle é feito através da aplicação de pequenos ângulos de rotação para as laterais. Esses ângulos são valores aleatórios entre -5 e 5 graus que são aplicados simultaneamente aos dois olhos em um intervalo de tempo aleatório entre 100 e 1000 milissegundos.

Reconhecimento de Tons DTMF para animação de movimentos faciais

A seguir é descrita a etapa de reconhecimento de tons DTMF para animações das feições faciais do humanóide virtual, representada por E440. O reconhecimento de tons DTMF é de conhecimento ordinário para uma pessoa versada na técnica, podendo ser implementado utilizando um filtro digital passa-faixa individual para cada freqüência. Quando um par de tons é reconhecido, o valor do número digitado é identificado, determinando qual gesto é desejado.

Para isto, são definidos alguns perfis de expressões que serão aplicados sempre que um comando correspondente for disparado. O perfil de animação pode ser relacionado a emoções, como por exemplo, felicidade, tristeza, raiva,

tédio, susto, confusão, sarcasmo, e pode também ser relacionado a movimentos da cabeça, por exemplo, sim e não, ou movimentos isolados do rosto, como mostrar a língua, levantar a sobrancelha, dentre outras. Assim, ao receber o tom DTMF referente a um número, o comando "fique feliz" é enviado para a etapa de fusão de informações, representado por E450.

Fusão de Informações

Nesta etapa, são avaliadas as prioridades entre os diferentes tipos de reconhecimento. Por exemplo, em relação aos movimentos labiais, o reconhecimento da vogal falada E420 tem prioridade na geração do gesto facial, isto é, ao receber um comando de mostrar a língua, o humanóide só mostrará a língua enquanto o usuário não falar. No momento que ele começar a falar, a movimentação da boca é feita apenas através das vogais.

Além disto, alguns dos perfis são temporários e outros são permanentes. Por exemplo, movimentos como sim e não são temporários, enquanto emoções como tristeza, alegria ou normal são permanentes. Os movimentos temporários possuem duração finita, ou seja, o *avatar* balança a cabeça indicando sim durante 5 segundos, voltando ao estado anterior. Desta forma, E450 realiza o controle de qual feição é realizada para que, em seguida, sejam gerados os

parâmetros FAP na etapa E460, que em seguida será animada na etapa E470.

Criação dos FAP

A etapa de formação dos parâmetros FAP E460 é descrita a seguir. A animação é baseada em um conjunto de pontos característicos da face ou *Feature Points* (FP). A codificação MPEG4 utiliza no total 84 FP, os quais, de acordo com a modalidade preferida da invenção, são usados um subconjunto de trinta e três pontos. Por exemplo, a codificação propõe dezoito FP para os lábios. Contudo, apenas oito pontos característicos podem animar eficientemente os lábios.

De acordo com a modalidade preferida da invenção, são utilizados apenas três FP para movimentação da cabeça, seis para a boca, quatro para cada olho, três para cada sobrancelha, cinco para o nariz, um para o queixo e dois para cada bochecha. Os principais FP são mostrados na figura 7.

Na figura 7, alguns FP são desenhados com uma bolinha cheia (701 - 726) e outros com uma bolinha vazia. Os primeiros são usados diretamente na animação, ou seja, são movimentados pelos FP em determinadas direções. Os segundos não são afetados pelos FP e permanecem imóveis durante todo o processo de animação. A função dos pontos estáticos é de

servir como limite para a deformação do rosto do humanóide quando um FP não estático é movimentado.

Por exemplo, ao receber o visema da vogal "O", a etapa E470 especifica os seis pontos extremos da boca para simular um círculo. Os FAP utilizam como base o deslocamento dos FP para modificação no formato geométrico do modelo facial.

A face neutra é definida a partir de um sistema de coordenas da mão direita (eixo X positivo para a direita, Y positivo para cima e Z positivo saindo do papel). Quando a face "olha" para a direção Z positiva, todos os músculos da face estão relaxados, pálpebras tangenciando a íris, pupilas medindo um terço de tamanho da pupila do modelo, lábios fechados formando uma linha horizontal de um canto ao outro da boca.

Diferentemente das soluções conhecidas, na presente invenção, os deslocamentos são sempre relativos aos FP na face neutra, e não em relação à posição anterior. Isto evita que a perda de um quadro de animação comprometa os quadros subseqüentes.

Cada perfil de animação facial é composto pelos índices do FP e dos deslocamentos de cada índice em relação à face neutra, um para cada eixo, dx, dy e dz. Por exemplo,

para fazer o *avatar* fechar os olhos, são utilizados quatro FPs e doze deslocamentos.

Além disto, em relação à boca, são utilizados apenas cinco possíveis configurações, uma para cada vogal e o perfil neutro (posição inicial). A partir dos perfis, o respectivo FAP é gerado e este é passado para a etapa de animação E470.

Animação Facial em Dispositivos Portáteis

Para a animação, a modificação de cada um dos FP faz com que outros pontos em torno do mesmo sejam afetados. Isso forma uma região de influência para cada um dos FP. Os pontos influenciados são calculados através de um método conhecido, em que o deslocamento de cada um pontos será dado por uma média ponderada dos deslocamentos dos seus FP influenciadores. Desta forma, a partir de todos os pontos mencionados, verifica-se o deslocamento de cada um destes em relação ao ponto atual. Quando a diferença é maior do que um limite de tolerância, os pontos do modelo são modificados, sendo possível animar as feições desejadas.

A presente invenção tendo sido descrita vai ser evidente para uma pessoa versada na técnica que muitas alterações e mudanças podem ser feitas na mesma, sem que se afaste do espírito ou do escopo da referida invenção, como definido nas reivindicações anexas.

REIVINDICAÇÕES

1) Método de síntese de movimento labial para animação de cabeças virtuais através do processamento de voz em dispositivos portáteis **caracterizado por** compreender as seguintes etapas:

- Configuração da comunicação para avaliação (E410) das opções pessoais do usuário e associação de um *avatar* entre um conjunto de humanóides virtuais disponíveis no dispositivo portátil;

- Aquisição de áudio recebido em formato padronizado;

- Análise de Energia do Áudio (E510) em uma quantidade fixa de amostras que formam um quadro de áudio, sendo a relação entre a energia do quadro atual e de seus anteriores calculada, dentro de uma janela de tempo para dimensionar o quanto a boca do *avatar* está aberta;

- Reconhecimento das vogais faladas (E420) por meio da análise dos quadros de áudio para reconhecer as vogais faladas, a partir da extração de características do sinal de áudio e a aplicação em uma rede neural progressiva com pesos fixos, gerados a partir de um treinamento feito fora do dispositivo portátil, utilizando um conjunto de amostras de áudio;

- Sintetizador de expressões faciais (E430) que gera artificialmente alguns gestos específicos do rosto que simulam movimentos naturais feitos por uma pessoa.

- Reconhecimento dos tons DTMF recebidos no canal de áudio (E440) para a animação da cabeça e gestos faciais;

- Fusão das informações reconhecidas e geradas artificialmente (E450) para formar uma única informação que será utilizada durante a animação do humanóide virtual, realizando uma análise de prioridades entre cada informação recebida, sendo que a vogal reconhecida tem prioridade maior do que as feições faciais reconhecidas pelos tons DTMF;

- Geração dos Parâmetros de Animação Facial através da conversão de instruções em parâmetros FAP (E460), utilizando as informações definidas na etapa anterior;

- Animação do humanóide virtual (E470) por meio da modificação das feições do humanóide através dos FAP recebidos.

2) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que a referida etapa de reconhecimento das vogais faladas (E420) compreende a geração do quadro de áudio (E500) que pode ser feita

através da amostragem da voz recebida pela linha telefônica (501) ou do arquivo disponível em uma mensagem MMS (502).

3) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 2, **caracterizado pelo** fato de que um quadro de áudio com duração de 32 ms é adquirido e para esta duração são processadas 512 amostras, que formam o quadro de áudio (503), sendo a conversão feita através do código implementado no DSP do dispositivo portátil.

4) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que após a referida etapa de cálculo da energia do sinal de voz em cada quadro adquirido (E510), cada valor é colocado em um *buffer* e é calculada a energia máxima nos últimos N ms ($\overline{E_{Max}}$) onde N varia entre 0,5 e 1,5 segundos, sendo preferencialmente utilizado o valor de 1 segundo.

5) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 4, **caracterizado pelo** fato de que o mínimo de energia para a detecção de voz $\overline{E_{Min}}$ é calculado como uma fração de $\overline{E_{Max}}$, a referida razão variando entre 1 % a 50 %, sendo preferencialmente utilizado o valor de 10%.

6) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 4, **caracterizado pelo** fato de que a razão entre a energia máxima e a energia do quadro atual é utilizada para controlar a dimensão de abertura da boca.

7) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que, além da energia E^j, são calculadas a energia total E^{tot} e a entropia residual *Wavelet* H^j, respectivamente descritas por:

$$E^{tot} = \sum_{j}^{J_{max}} E^j \quad , \tag{1}$$

$$H^j = -\frac{E^j}{E^{tot}} \cdot \log \frac{E^j}{E^{tot}} \tag{2}$$

onde a energia é calculada para níveis específicos da decomposição e as características podem ser calculada em qualquer combinação entre os níveis da transformada de 1 a J_{max}, preferencialmente são utilizados os níveis 1, 3, 4 e 5 para calcular os coeficientes, sendo a escolha de não utilizar um nível específico feita através de testes experimentais com várias combinações de características

para verificar a combinação que apresenta melhor desempenho.

8) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 7, **caracterizado pelo** fato de que, após o cálculo de E^{tot} e H^{j}, na etapa E630, as características são aplicadas em uma rede neural progressiva, que possui, preferencialmente, N neurônios na camada de entrada (N = número de características usadas), oito na camada escondida e seis na camada de saída, em que as cinco primeiras saídas indicam cada tipo de vogal e a sexta saída indica a ausência de vogal.

9) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que o controle dos movimentos do globo ocular é feito através da aplicação de pequenos ângulos de rotação para as laterais, tendo os referidos ângulos valores aleatórios entre -5 e 5 graus que são aplicados simultaneamente aos dois olhos em um intervalo de tempo aleatório entre 100 e 1000 milissegundos.

10) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que na referida etapa de

fusão de informação alguns dos perfis são temporários e outros são permanentes.

11) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que a referida etapa de formação dos parâmetros FAP (E460) é baseada em um conjunto de pontos característicos da face ou *Feature Points* (FP).

12) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 11, **caracterizado pelo** fato de que alguns dos referidos pontos característicos da face são usados diretamente na animação em determinadas direções.

13) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que os deslocamentos são sempre relativos aos FP na face neutra e não em relação à posição anterior, evitando que a perda de um quadro de animação comprometa os quadros subseqüentes.

14) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que cada perfil de animação facial é composto pelos índices do FP e dos deslocamentos de cada índice em relação à face neutra, um para cada eixo, dx, dy e dz.

15) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que, em relação à boca, são utilizados apenas cinco possíveis configurações, uma para cada vogal e o perfil neutro (posição inicial), sendo que a partir dos perfis, o respectivo FAP é gerado e este é passado para a etapa de animação E470.

16) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que, para a animação, a modificação de cada um dos FP faz com que outros pontos em torno do mesmo sejam afetados, formando uma região de influência para cada um dos FP, em que os pontos influenciados são calculados através do deslocamento de cada um pontos será dado por uma média ponderada dos deslocamentos dos seus FP influenciadores, sendo a verificação do deslocamento de cada um destes em relação ao ponto atual feita a partir de todos os pontos mencionados.

17) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que permite a animação dos movimentos da cabeça, a seleção de feições e o reconhecimento do movimento labial.

18) Método de síntese de movimento labial para animação de cabeças virtuais, de acordo com a reivindicação 1, **caracterizado pelo** fato de que é adaptado para a extração de características para o reconhecimento das vogais faladas em diferentes entonações e sotaques, particularmente do português.

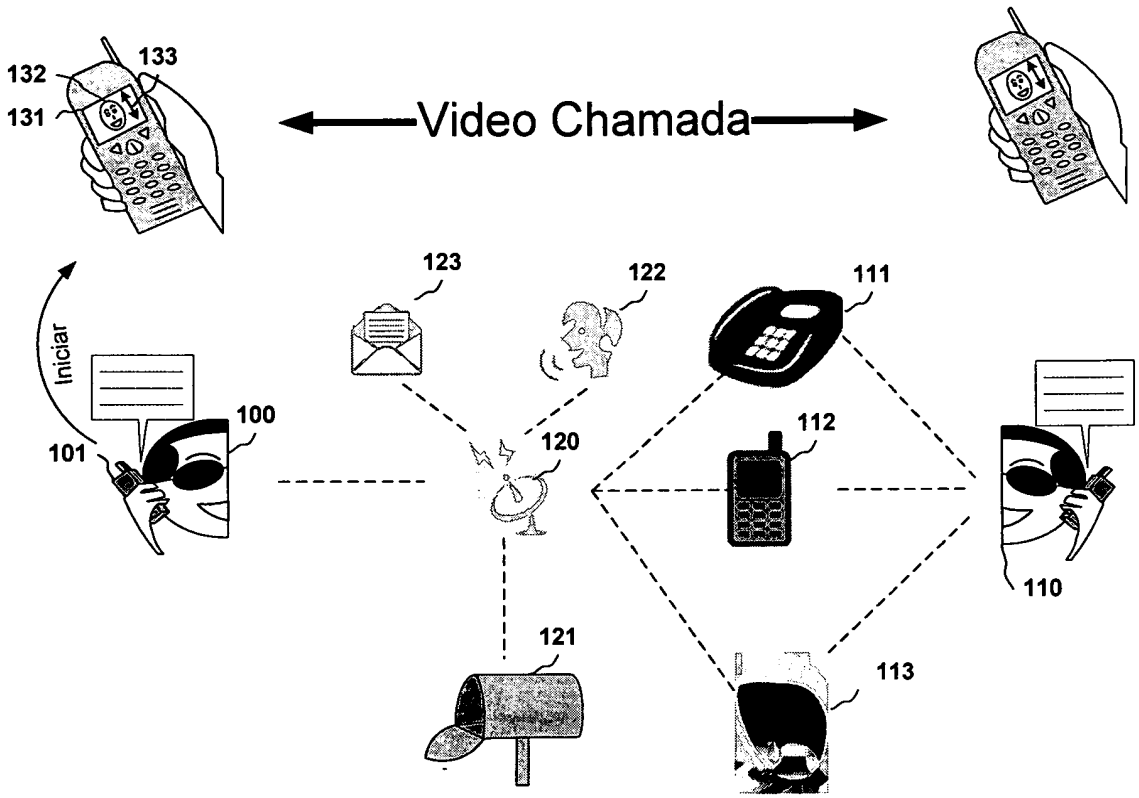

132
133
131
Video Chamada
Iniciar
123
122
111
100
101
120
112
110
121
113

FIG 1

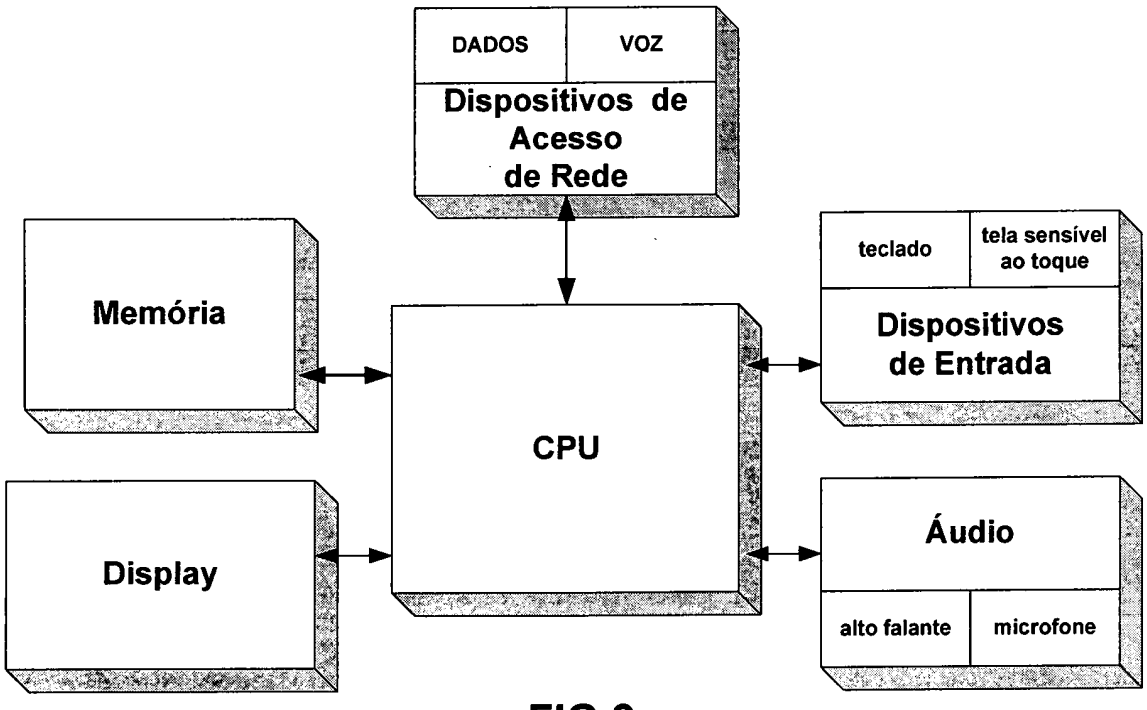

DADOS
VOZ
Dispositivos de Acesso de Rede
Memória
CPU
teclado
tela sensível ao toque
Dispositivos de Entrada
Display
Áudio
alto falante
microfone

FIG 2

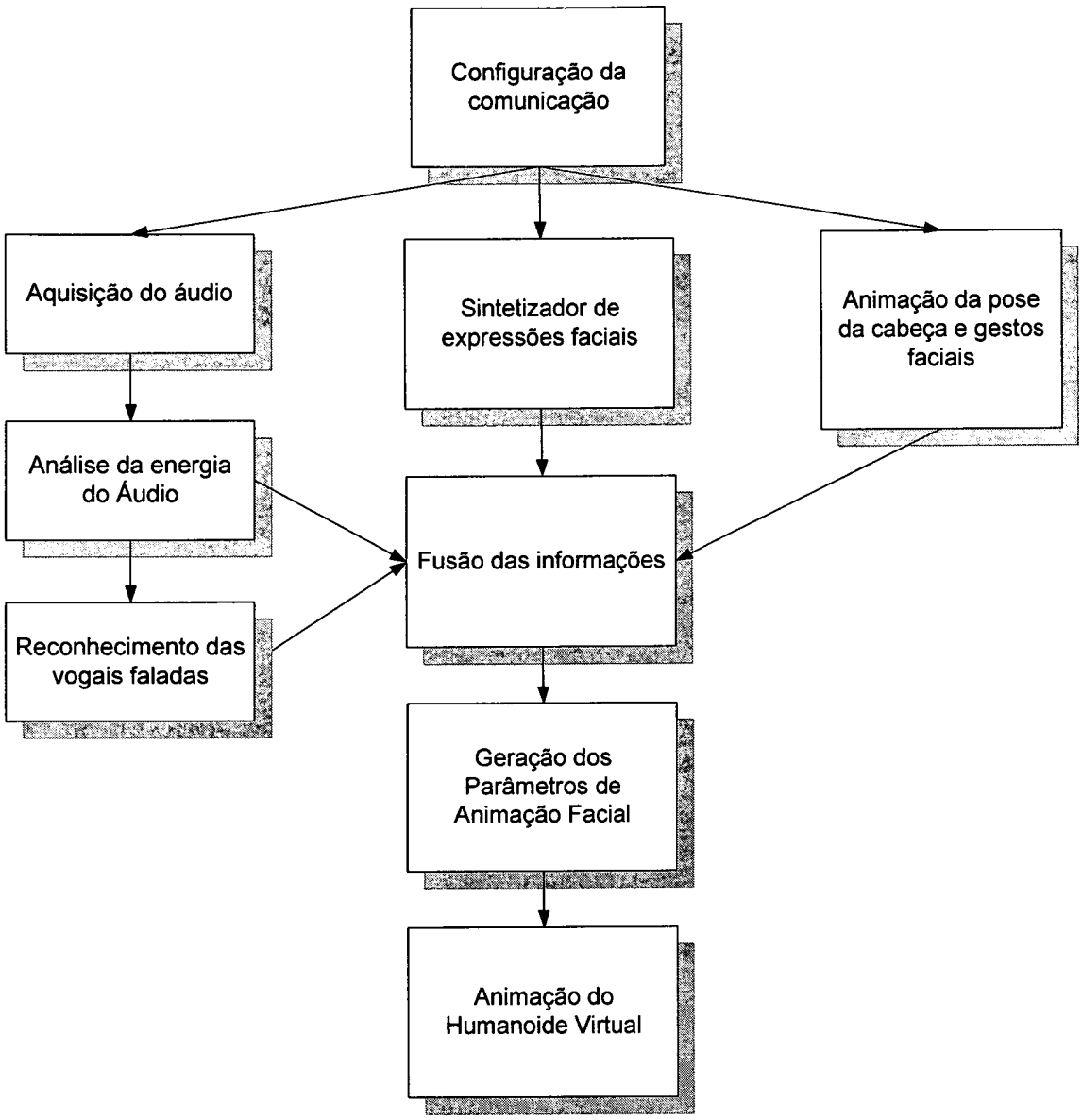

Configuração da comunicação
Aquisição do áudio
Sintetizador de expressões faciais
Animação da pose da cabeça e gestos faciais
Análise da energia do Áudio
Fusão das informações
Reconhecimento das vogais faladas
Geração dos Parâmetros de Animação Facial
Animação do Humanoide Virtual

FIG 3

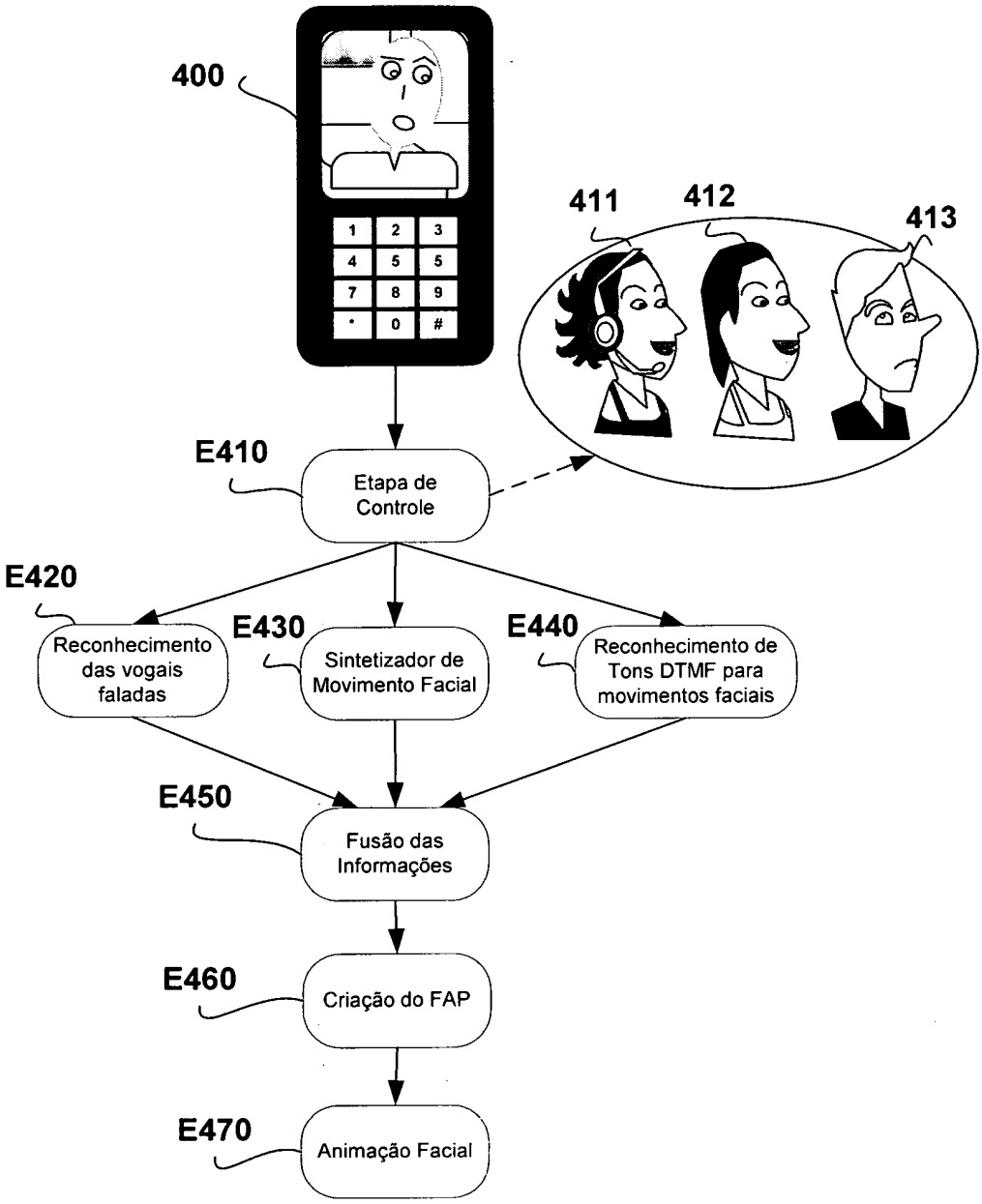

400
411
412
413
1 2 3
4 5 5
7 8 9
* 0 #
E410
Etapa de Controle
E420
Reconhecimento das vogais faladas
E430
Sintetizador de Movimento Facial
E440
Reconhecimento de Tons DTMF para movimentos faciais
E450
Fusão das Informações
E460
Criação do FAP
E470
Animação Facial

FIG 4

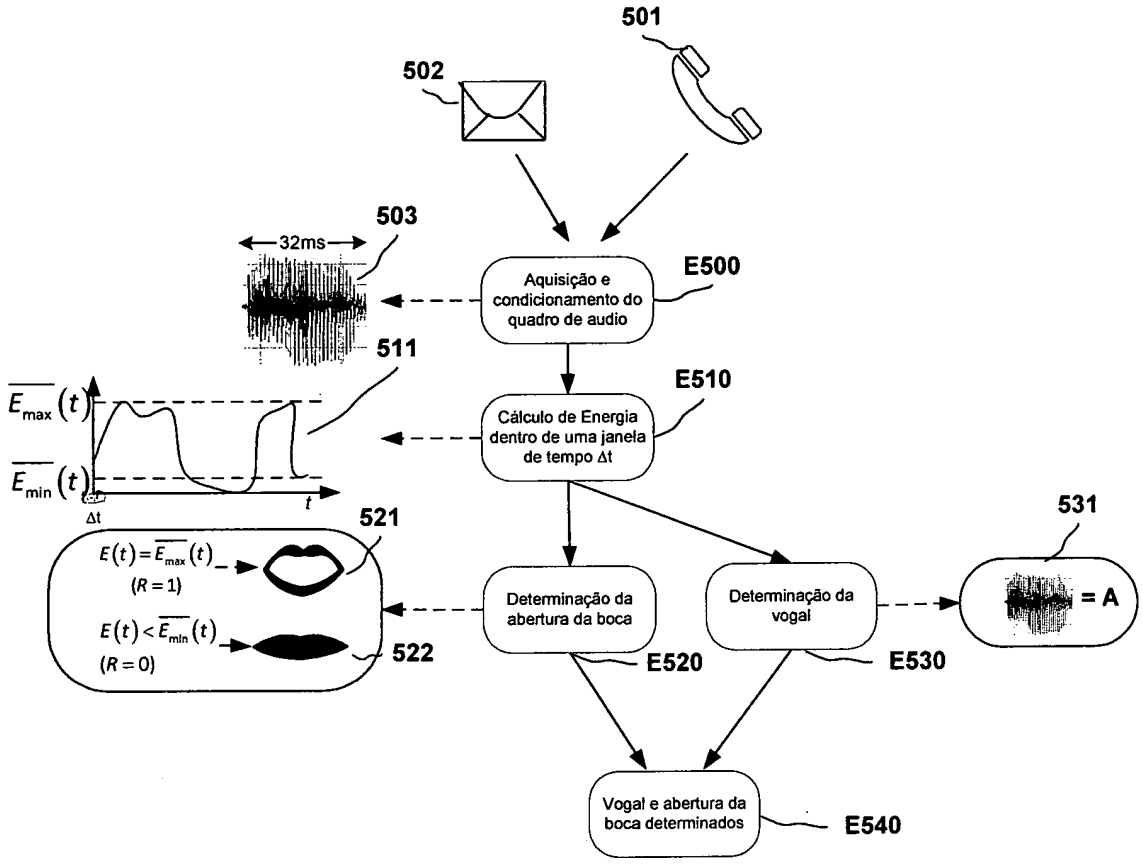

501
502
503
32ms
Aquisição e condicionamento do quadro de audio
E500
511
$\overline{E_{max}}(t)$
$\overline{E_{min}}(t)$
Δt
t
Cálculo de Energia dentro de uma janela de tempo Δt
E510
521
$E(t)=\overline{E_{max}}(t)$
(R = 1)
$E(t)<\overline{E_{min}}(t)$
(R = 0)
522
Determinação da abertura da boca
E520
Determinação da vogal
E530
531
= A
Vogal e abertura da boca determinados
E540

FIG 5

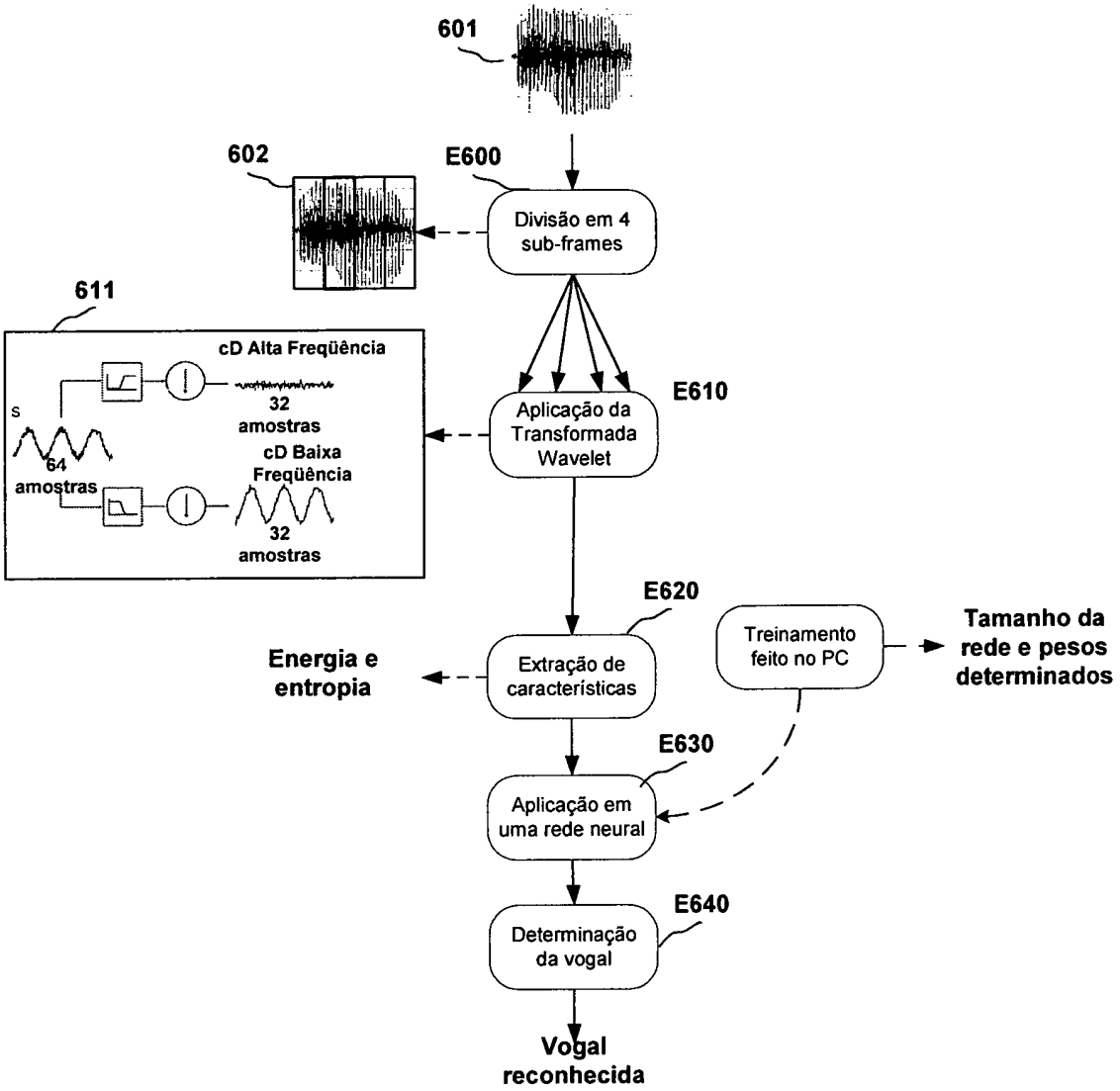

601
602
E600
Divisão em 4 sub-frames
611
cD Alta Freqüência
32 amostras
S
64 amostras
cD Baixa Freqüência
32 amostras
E610
Aplicação da Transformada Wavelet
E620
Energia e entropia
Extração de características
Treinamento feito no PC
Tamanho da rede e pesos determinados
E630
Aplicação em uma rede neural
E640
Determinação da vogal
Vogal reconhecida

FIG 6

PI0905940-6

Resumo da Patente de Invenção para: **"MÉTODO DE SÍNTESE DE MOVIMENTO LABIAL PARA ANIMAÇÃO DE CABEÇAS VIRTUAIS ATRAVÉS DO PROCESSAMENTO DE VOZ EM DISPOSITIVOS PORTÁTEIS"**.

Aspectos da presente invenção especificam um método de comunicação utilizando um humanóide virtual animado durante chamadas convencionais. De acordo com a presente invenção, a animação é feita utilizando um sistema de reconhecimento das vogais faladas para animação dos lábios, associado ao reconhecimento de tons DTMF para animação dos movimentos da cabeça e feições faciais. Esse sistema se diferencia dos sistemas conhecidos por usar técnicas computacionalmente eficientes. Isto implica em um trabalho de otimização e ajuste de técnicas de processamento digital de sinais para viabilizar sua implementação em dispositivos portáteis.

O método aqui descrito pode ser implementado de forma genérica em dispositivos portáteis como PDAs, celulares e *Smart Phones* que possuam acesso ao serviço de telefonia móvel.