Interação Multimodal em Aplicações Hipermídia

Flávio Miranda de Farias, Allysson Chagas Carapeços, Débora Christina Muchaluat Saade

MídiaCom - Instituto de Computação – Universidade Federal Fluminense (UFF)

fmflavio@gmail.com, allyssoncc@gmail.com, deboracms@midiacom.uff.br

Resumo. *Cada vez há mais a utilização da tecnologia no dia a dia, até mesmo em coisas comuns, ela está presente. A digitalização de uma mídia (áudio, imagem, vídeo etc.) pode parecer algo simples para o usuário, mas exige muito processamento para que seja exibida de forma transparente. Um termo da tecnologia que tem tido cada vez mais relevância é a interação multimodal, modalidade da área de IHC onde se estuda a maneira em que o ser humano consegue interagir com a máquina. Por se tratar de interação, tem ligação direta com os sentidos humanos, muitas vezes combinando mais de um deles. Quando uma aplicação multimídia permite interação, denomina-se hipermídia, que envolve a integração de vários tipos de mídia onde o usuário pode navegar pelos dados como achar conveniente. Neste trabalho serão apresentados conceitos e estado da arte sobre interação multimodal e hipermídia, bem como a utilização de linguagens de autoria que auxiliam neste processo.*

1. Introdução

Interação multimodal é uma modalidade da área de Interação Humano-Computador (IHC) quem vem ganhando destaque e amadurecendo com a evolução das técnicas de usabilidade de sistemas (ROGERS; SHARP; PREECE, 2013). Enquanto as aplicações hipermídia, já consolidadas pelos criadores e bem aceitas pelos usuários por diversos fatores, normalmente utilizam formas interativas comumente vendidas como padrões influenciados pelo mercado comercial, não é comum encontrar muitas aplicações que ofereçam interface multimodal, mas a seguir, serão mostradas algumas tecnologias relacionadas.

Os principais dispositivos de entrada e modos de interação em aplicações atuais são (CHURCHILL, 2018; LAZAR; FENG; HOCHHEISER, 2017; ROGERS; SHARP; PREECE, 2013):

- Teclado e Mouse (*touchpad*): normalmente utilizada por usuários de microcomputadores, laptops e similares. Este kit é capaz de atender bem a navegabilidade através do click em links e botões, além de oferecer atalhos rápidos pelo teclado. Não é sempre intuitivo, além de não oferecer conforto em todos os ambientes de uso como um sofá, poltrona ou até mesmo em pé.
- Telas sensíveis ao toque: bem intuitiva principalmente para usuários jovens. Essa forma de interação é muito boa para pequenos aparelhos, não sendo adequada para grandes telas ou televisores. Existem nos formatos de tecnologia:

    - Resistiva, onde são necessários cliques na tela com maior pressão para os cliques serem registrados, além de não darem suporte ao multitoque. Normalmente apresentados em aplicações de muito uso como caixas eletrônicos, celulares e tablets antigos; e
    - Capacitiva onde utiliza parte da carga elétrica contida nos dedos para afetar a corrente presente nas telas, e com isso registrar o clique. Permitem o multitoque, e são usados em grande parte dos dispositivos atuais, como smartphones e tablets.

- Comandos por voz: atualmente, mais eficientes e precisos, devido a tecnologia de Aprendizado de Máquina pertencente a área de Inteligência Artificial (IA). Essa tecnologia está sendo empregada cada vez mais por sistemas operacionais como Microsoft Windows, iOS e Android no formato de assistente pessoal, ou assistente de acessibilidade a pessoas com deficiência, além de equipamentos que são assistentes residências. Não é comum sua utilização para navegar em aplicações hipermídia. Possui ainda limitações em relação a sotaques e linguagens estrangeiras.
- Comando por Gestos: para o uso dessa tecnologia, há diversas formas de atuar, mas normalmente utiliza sensores infravermelhos, sensores acústicos ou câmeras. Normalmente aplicada a jogos eletrônicos, não é muito empregada em aplicações hipermídia. Necessita de aperfeiçoamento, pois assim como os Comandos por Voz, comandos por gestos também têm problemas com reconhecimento de padrões de comandos, como no caso de gestos errados ou iluminação inadequada. Usuários normalmente relatam cansaço em restos repetitivos e bruscos. Pode trabalhar com reconhecimento de gestos das mãos, ou outras partes do corpo, por exemplo olhos, ou até o corpo todo, além de reconhecimento de outros objetos.
- Controle remoto: item comum presente na casa da população por ser item de controle de televisores e outros eletroeletrônicos. É encontrado de diversas formas, com diversas tecnologias como: infravermelho, *bluetooth* e Wi-Fi. Pode servir como apontador, seguindo a ideia de um mouse ou simplesmente enviando comandos através de teclas.
- Comando por ondas cerebrais: o cérebro trabalha com ondas cerebrais que podem ser capturadas por meio de eletrodos ou capacetes e interpretadas como comandos, esses comandos podem ser enviados para uma interface e aceitos como controle. Atualmente essa tecnologia ainda está amadurecendo não sendo empregada comercialmente, além de ser necessário um treinamento do usuário para utilização (RAMADAN; VASILAKOS, 2017).
- Giroscópio e acelerômetro: normalmente embutidos em aparelhos da telefonia móvel, fornecem informações de deslocamento no eixo x, y e z, além da aceleração da posição inicial até a final em um determinado tempo. Com aparelhos que possuem essa tecnologia, pode ser atribuído controle normalmente em 3D, normalmente utilizados em jogos eletrônicos ou controles de drones.
- Controles e *joystick*: a indústria de jogos eletrônicos é a indústria mais aberta a adoção de tecnologias. Suas plataformas de jogos adotam controles e joysticks de modelos tradicionais até os mais elaborados, contendo várias das tecnologias

anteriormente citadas, porém seus controles são normalmente caros em relação às demais tecnologias até por embarcar diversas delas. Usualmente não são usados para navegação em aplicações hipermídia.

- Telas secundárias: tecnologia que anda sendo aderida discretamente. Trata-se da utilização de outras telas além da principal, tanto para controle básico de equipamentos multimídia, quanto para jogos eletrônicos, pode atuar como extensão da tela principal em algumas aplicações. A comunicação usualmente utiliza como base a rede sem fio local do usuário, mas sofre com latência em aplicações online.

Como descrito anteriormente, há diversas maneiras de interagir com aplicações em geral, sendo até utilizadas diversas categorias de forma simultânea, mas de forma geral, normalmente se utilizam métodos padronizados para cada tipo de equipamento que normalmente são:

- Televisores e players multimídia: controle remoto;
- Computadores: teclado e mouse (*touchpad*);
- Videogames: controles e *joystick*;
- Tablets e smartphone: tela sensível ao toque.

O objetivo principal deste trabalho é abordar de forma teórica algumas tecnologias de interação multimodal com foco em tecnologias hipermídia. Além disto, citar contexto histórico, de mercado e de usabilidade, citando equipamentos, modos de interação e plataformas que podem adotar, como por exemplo, a TV digital.

Este trabalho abordará na Seção 2 os conceitos sobre interação multimodal em aplicações multimídia. Na Seção 3, serão apresentados trabalhos relacionados ao estado da arte do tema. Na Seção 4, o foco será discutido nos trabalhos relacionados a hipermídia, complementando na Seção 5 com linguagem de autoria hipermídia. Para finalizar o estudo, apresentam-se as conclusões na Seção 6.

2. Interação multimodal em aplicações multimídia

Segundo o livro (PRESSMAN, 2011), softwares da categoria de aplicação, são programas independentes que solucionam uma necessidade específica de negócio, ou seja, são softwares que atendem uma exclusiva finalidade.

As aplicações podem ser controladas através de controle explícito ou oculto, de acordo com seu *layout* (ROGERS; SHARP; PREECE, 2013):

- Controle explícito – quando a aplicação mostra ao usuário quais controles ele tem e pode utilizar, normalmente exibido no formato de menus, botões, *labels* ou descrições vocais.
- Controle oculto – quando a aplicação não informa como ser controlada, por exemplo, a forma como controlamos através de controle remoto.

Na era digital muitas tecnologias surgiram para dar controle das aplicações ao usuário. Com o tempo, várias dessas tecnologias foram se unindo ou completando, com isso, surge a terminologia multimodal, que vem da comunicação na qual coexistem diversas

modalidades comunicativas (fala, gestos, texto, processamento de imagem, etc.), ou seja, muitas modalidades (GUEDES; AZEVEDO; BARBOSA, 2017; MATTOS; MUCHALUAT-SAADE, 2018).

As pessoas se comunicam de forma multimodal no dia a dia, por exemplo, quando há conversa, é possível entender a fala, as expressões do rosto e até expressões gestuais, com isso, sendo possível receber inúmeras informações em paralelo que serão agrupadas para formar uma ou mais informações, sendo que se fossem separadas poderiam corromper o significado da comunicação (TALARICO NETO, 2011). Além disso, Interfaces multimodais devem adaptar-se às necessidades e habilidades de diferentes usuários, bem como diferentes contextos de uso. A adaptabilidade dinâmica permite que a interface seja degradada normalmente, aproveitando modalidades complementares e complementos de acordo com as mudanças na tarefa e no contexto (REEVES et al., 2004).

Em (TURK, 2014), conceituam-se interfaces multimodais como sistemas interativos que buscam alavancar recursos humanos naturais para comunicação via fala, gesto, toque, expressão facial e outras modalidades, trazendo métodos mais sofisticados de reconhecimento e classificação de padrões para a interação humano-computador. Além disso, Turk lista na Tabela 1 as modalidades relevantes que considera.

Tabela 1. Modalidades sensoriais humanas relevantes para a interação multimodal (Adaptado de (TURK, 2014)).

Modalidade	Exemplo
Visual	Localização da face
	Olhar
	Expressão facial
	Leitura labial
	Identidade baseada em rosto (e outras como idade, sexo, raça, etc.)
	Gesto (cabeça / face, mãos, corpo)
	Linguagem de sinais
Auditivo	Entrada de fala
	Áudio sem fala
Tocar	Pressão
	Localização e seleção
Outros sentidos	Gesto
	Captura de movimento baseada em sensor

No início da década de 80, o trabalho de (BOLT et al., 1980) apresentou uma demonstração multimodal, antes mesmo de o termo existir, no qual o usuário poderia realizar comandos de voz e apontamentos em uma interface gráfica rudimentar preparada para o experimento. Interessante que na época não existiam as interfaces gráficas hoje conhecidas em nossos sistemas operacionais e o autor, além de se preocupar em ter de

construir a aplicação, também tinha de se preocupar com os controles e até o ambiente de uso, como ilustrado na Figura 1.

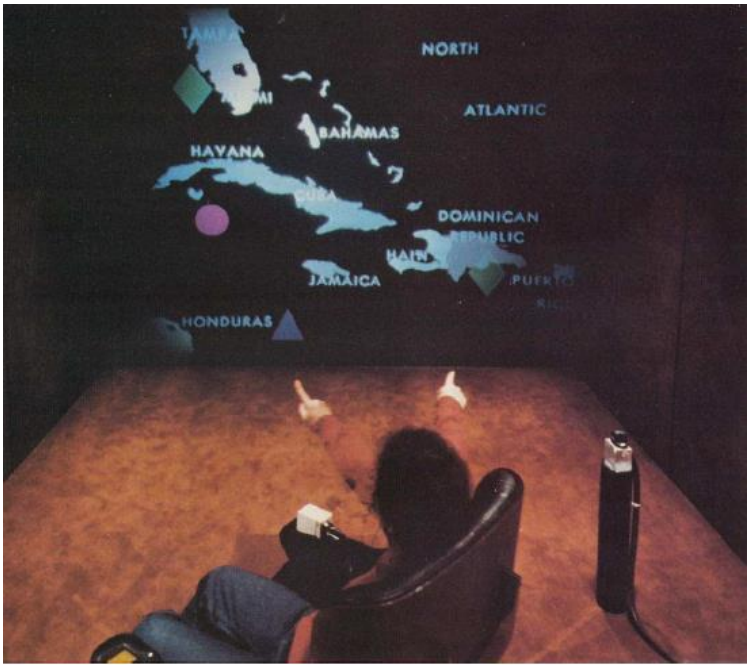

Figura 1. Conversando e apontando para itens na tela da sala de mídia (BOLT et al., 1980).

Após o trabalho de (BOLT et al., 1980), a tecnologia não evoluiu muito no meio multimodal, provavelmente porque em nível computacional ainda não se tornava viável a produção, e enquanto não havia um apelo comercial para adoção, era mais simples e viável oferecer controles individuais.

Contudo, em 2006, a Nintendo lançou o videogame Nintendo Wii, que como novidade e diferencial, apresentava um controle com formato de controle remoto, apontador infravermelho, acelerômetro e giroscópio, permitindo uma maneira diferente para interagir com os jogos, isso foi uma novidade, impulsionando muitas vendas do console e liderança no mercado de videogames. Este controle, denominado de Wii Remote, possibilitava a interação por movimentos, possibilitando a implementação de vários jogos de esporte e interações inéditos, dando início a uma corrida comercial por novas tecnologias de interação (ALVES et al., 2018; SOUSA, 2011).

Posteriormente, após os controles do Wii, academicamente muito se buscou a utilização dos recursos dos controles para fazer experimentos, mas a tecnologia ainda era precária em relação aos smartphones de hoje, por exemplo, não apresentando precisão em seus movimentos ou apontamentos. Com isso, a Microsoft em 2010, apresentou o Kinect, anteriormente chamado “projeto natal”, por ser idealizado e desenvolvido por um brasileiro na cidade de Natal. O Kinect é um acessório do videogame Xbox 360 e possibilitava finalmente interações multimodais com jogos. Nesse mesmo ano, a Sony lançou o *PlayStation Move*, com tecnologia similar a do Nintendo Wii, porém com maior precisão. A tecnologia da Sony é composta de um controle com uma esfera luminosa, e necessita de uma câmera (*PlayStation Eye*) que faz a captura de voz, imagens e movimentos do controle e corporais (ALVES et al., 2018; SCHLÖMER et al., 2008).

O Kinect possui duas câmeras de baixa qualidade, sendo a primeira uma RGB possibilitando reconhecimento de face, e a segunda uma de profundidade usada para construção do esqueleto dos usuários, além de um microfone para reconhecimento de comandos de voz. A Microsoft tratou de disponibilizar com o tempo, como estratégia de mercado, uma versão para PC com Windows e Kits para desenvolvedores grátis, com isso houve muitos trabalhos acadêmicos de diversas finalidades (ALVES et al., 2018).

O mercado a partir de então esteve recebendo cada vez mais trabalhos e pesquisa na área de IHC, como reconhecimento de gestos, tela secundária, reconhecimento objetos, *traking* (rastreamento), posicionamento e reconstrução 3D de movimentação de objetos, reconhecimento de voz como comandos ou complemento de controle, sensores biométricos e muitas outras formas de pesquisa, possibilitando as interações multimodais atuais.

3. Trabalhos relacionados

Em (DUMAS; LALANNE; OVIATT, 2009) , os autores do artigo abordam sobre *Multimodal User Interfaces* (MUI). Este trabalho é muito interessante visto que se preocupa em destacar a importância da autoria multimídia voltada a interação com usuário. Usualmente a área que trabalha com a interação do usuário é a IHC (ROGERS; SHARP; PREECE, 2013).

Guedes em *Extending multimedia languages to support multimodal user interactions* (GUEDES; AZEVEDO; BARBOSA, 2017), propõe através de diversas tecnologias estender a linguagem multimídias como NCL a prover suporte a vários usuários de forma multimodal. O autor relata que há limitações nestas linguagem por terem foco em mídias como texto, imagem, áudio e vídeo, porém, a comunidade que trabalha com MUIs tem provido extensões que tem contribuído muito com o estado da arte.

Outro trabalho que aborda autoria multimodal é (DUMAS; LALANNE; OVIATT, 2009). Os autores propuseram uma arquitetura abstrata para sistemas multimodais. Relatam que cada interação entre usuário e sistema em um MUI pode ser vista como uma ação de percepção, denominado laço de interação multimodal, entre o sistema multimodal e o usuário. No ambiente de controle, o usuário executa ações através de atividades de comunicação humana (por exemplo, fala, gestos, toque) e percebe as ações do sistema (resposta) através de seus sentidos (por exemplo, visão, audição). O sistema percebe atividades humanas por meio de dispositivos de entrada e sensores (por exemplo, teclado, câmera de vídeo e microfone), reconhece as ações esperadas do usuário e realiza ações por intermédio de dispositivos de saída audiovisuais e dispositivos atuadores (por exemplo, monitor, alto-falantes, háptico).

De forma contextual, TURK (2014) explica que múltiplos sentidos, sequencialmente e em paralelo, podem explorar passivamente e ativamente nosso ambiente, para confirmar as expectativas sobre o mundo e para perceber novas informações. Demonstra que é possível experimentar estímulos externos através da visão, audição, tato e olfato, e sentimos respostas musculares como reações reflexivas de nosso sistema nervoso musculatório. Além disso, revela que uma dada modalidade de sensoriamento pode ser usada para estimar simultaneamente várias propriedades úteis do ambiente - por exemplo, formas de áudio podem ser usadas para determinar a identidade

e localização de um falante, para reconhecer as palavras do interlocutor e interpretar a prosódia do enunciado, para estimar o tamanho e outras características do espaço físico circundante, e identificar outras características do ambiente e atividades periféricas simultâneas.

Um artigo bem completo sobre este assunto, porém com foco em controle dos menus do sistema operacional de *Smart* TVs é o trabalho (LEE; KAOLI; HUANG, 2014). Este trabalho foi escrito em um período em que as *Smart* TVs começaram a se popularizar globalmente e as fabricantes de televisores, tentavam inovar oferecendo diferenças, inclusive interação multimodal utilizando captura de gestos e até áudio, incluindo câmeras e microfones em TVs de alto padrão como ilustrado na Figura 2.

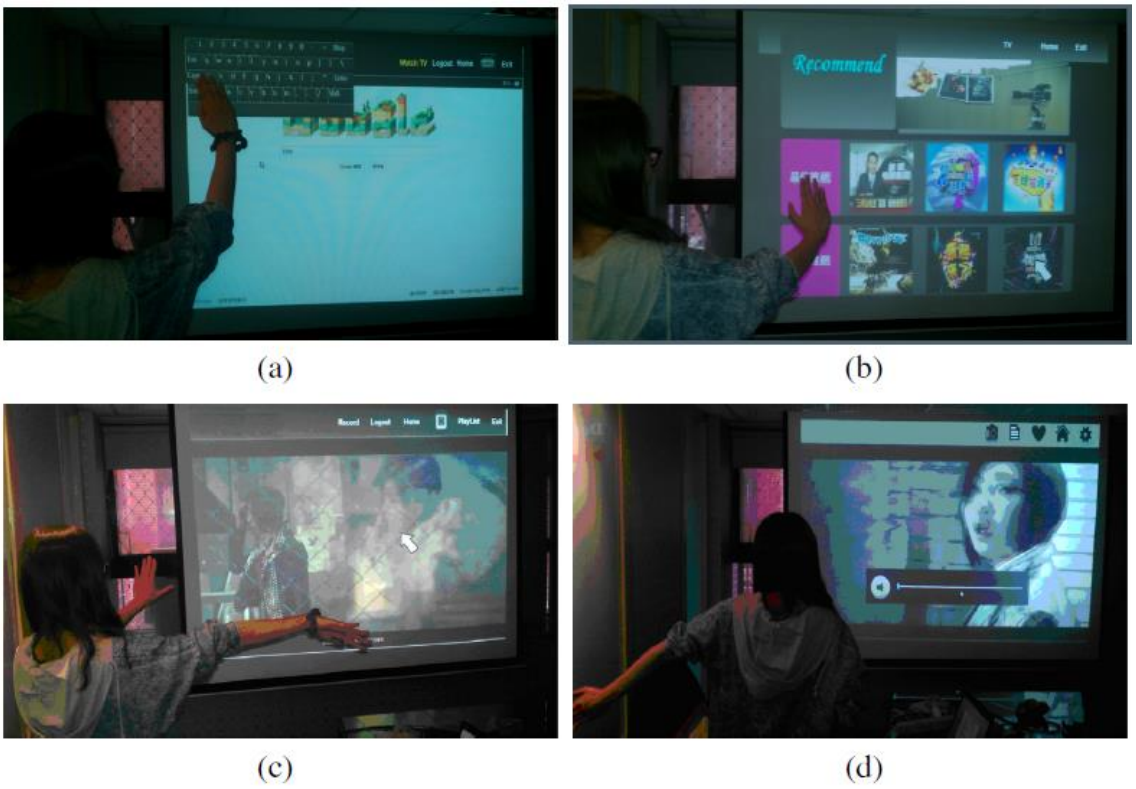

Figura 2. Controle por gestos corporais (LEE; KAOLI; HUANG, 2014).

Uma proposta que pode ser adaptada ao contexto multimodal é de (MAKLEYSTON GOMES PEREIRA; JOSÉ DA SILVA SILVA; LÍVIO VASCONCELOS GUEDES, 2017). Os televisores com suporte ao sistema brasileiro de televisão digital (SBTD) possuem entre outras coisas, suporte a aplicações hipermídia, e neste trabalho, os autores abordaram a tecnologia "internet das coisas" (IoT), neste trabalho os autores expandiram a função da TV há um controle central de equipamentos diversificados, ou seja, um *smart hub* multimídia.

Ainda no tema SBTVD, o artigo (GOMES SOARES; MORENO; GUEDES VASCONCELOS, 2015a) apresenta um modelo de controle hierárquico que visa apoiar linguagens declarativas direcionadas a aplicativos de Web e de TV digital. O modelo tem seu uso potencial em algumas linguagens de hipermídia declarativas padrão (baseadas em HTML, SMIL e NCL). Neste trabalho, implementam algumas formas de interação multimodal utilizando Kinect, *multi-device* entre outras coisas (ver Figura 3), porém eles

citam que a implementação ainda enfrenta grandes problemáticas como as limitações de captura do Kinect.

Figura 3. Controle por gestos utilizando Kinect (GOMES SOARES; MORENO; GUEDES VASCONCELOS, 2015a)

Como ferramenta de autoria hipermídia, o trabalho *STEVE: Spatial-Temporal View Editor for Authoring Hypermedia Documents* (MATTOS; MUCHALUAT SAADE, 2016), utiliza o paradigma da linha de tempo para usuários sem conhecimento de programação, a aplicação baseada em modelo de documento hipermídia, utiliza e aborda os eventos em um editor gráfico, para edição de visão espacial-temporal de um documento, além disso, o STEVE exporta aplicativos hipermídia para documentos NCL e HTML5 para realizar diferentes plataformas de execução (ver Figura 4). A ferramenta de autoria, possui base para suportar a atualizações futuras e dar suporte a interações multimodais. Em outro trabalho relacionado *MultiSEM: A Mulsemedia Model for Supporting the Development of Authoring Tools* (MATTOS; MUCHALUAT-SAADE, 2018), os autores estendem as funcionalidade da ferramenta ao suportar atuadores multissensoriais propondo o conceito, mulsemídia (MulSeMedia - Multiple Sensorial Media).

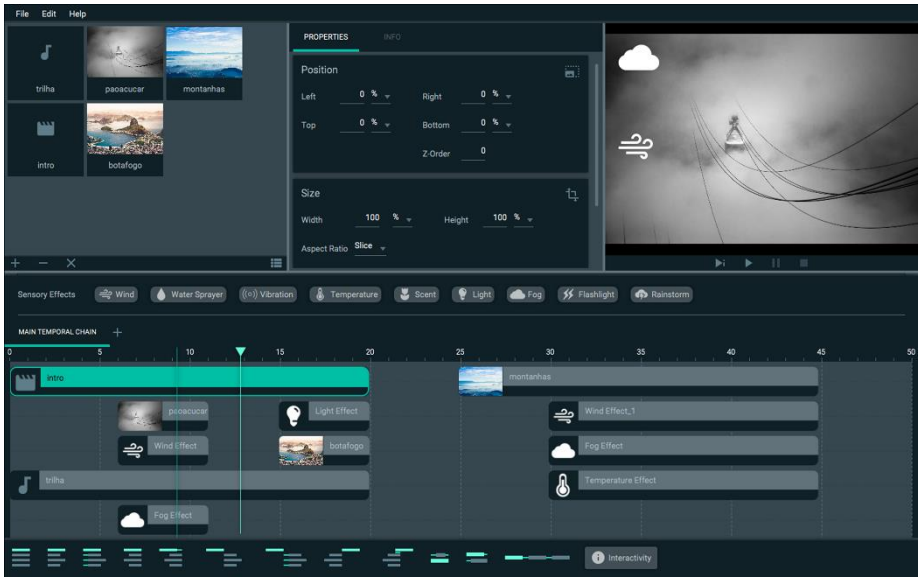

Figura 4. Interface do STEVE 2.0 com efeitos sensoriais (GOMES SOARES; MORENO; GUEDES VASCONCELOS, 2015b).

Apesar de diversos trabalhos que abordam bem tecnologias de interação multimodal em equipamentos e aplicações multimídia, jogos, TVs e afins, essa abordagem, não é bem explorada ainda em aplicações hipermídia.

4. Linguagens de Autoria Declarativas para Interação Multimodal

As linguagens de autoria declarativas específicas para desenvolvimento de uma aplicação multimodal permitem que tenha abstração dos detalhes de um código. Com a abstração, o desenvolvimento se torna mais ágil, economizando tempo e esforço. É importante que a linguagem possibilite o reúso e portabilidade. Nesta seção iremos apresentar algumas linguagens de autoria declarativas utilizadas para aplicação multimodal.

4.1. VoiceXML

No ano de 1995, por meio de reuniões informais, Dave Ladd, Chris Ramming, Ken Rehor, and Curt Tuckey da AT&T Research iniciaram a discussão sobre uma linguagem de marcação que fornecesse serviços da web para telefones comuns. Inicialmente este projeto se chamou AT&T Phone Web. Já em 1999, a AT&T e Lucent possuíam dialetos incompatíveis da linguagem *Phone Markup Language* (PML), a IBM tinha o SpeechML e a Motorola tinha o VoxML. Coma esta diversidade de linguagens, surgiu a necessidade de padronização, criando-se assim o Fórum do VoiceXML, composto inicialmente por AT&T, Lucent e Motorola.

A primeira versão produzida pelo Fórum foi o VoiceXML 0.9, que foi a combinação das tecnologias propostas anteriormente, em conjunto com novos recursos, entre eles o *Dual Tone Multi Frequeny* (DTMF), que são os sons emitido ao pressionar as teclas do telefone. Após a publicação, a comunidade cresceu, recebendo uma grande quantidade de contribuições que geraram melhorias na linguagem. Com as contribuições, o Fórum lançou o VoiceXML 1.0, em março de 2000, sendo enviado no mês seguinte para avaliação do World Wide Web Consortium (W3C).

Após a aceitação da versão 1.0 pelo W3C, foram feitas modificações para correção de erros na especificação, desenvolvimento de novos padrões e gramáticas de reconhecimento de fala e marcação de texto para fala, surgindo assim a versão 2.0. Esta versão da linguagem passou a ser recomendada em março de 2004 pela W3C (W3C, 2018). A Tabela 2 apresenta alguns elementos da liguagem VoiceXML separados por escopo.

Tabela 2. Principais elementos do VoiceXML (adaptado de (TALARICO NETO, 2011))

Escopo	Elementos (tags)
Variáveis e propriedades	<assign>, <meta>, <param>, <property>, <script>, <value>, <var>, <clear>
Síntese de fala e saída de áudio	<audio>, <block>, <enumerate>, <reprompt>, <sayas>, <break>
Tratamento de erros	<catch>, <throw>, <error>, <help>, <noimput>, <nomatch>
Fluxo do diálogo	<choice>, <elseif>, <form>, <goto>, <exit>, <subdialog>, <submit>, <else>, <link>, <return>, <menu>, <option>, <if>
Entradas de usuários e gramáticas	<field>, <dtmf>, <record>, <filled>, <grammar>
Integração com sistemas de telefonia	<transfer>, <disconnect>
Definição de documento	<vxml>, <initial>, <object>

Uma *minor release* foi recomendada em junho de 2007 pela W3C, a versão 2.1, onde surgiram 2 novos elementos (<data> e <foreach>), e foram aprimorados outros 6 (<disconnect>, <grammar>, <mark>, <property>, <script> e <transfer>) (OSHRY et al., 2018). A versão 3.0 do VoiceXML está em fase de implementação, tendo um rascunho publicada em dezembro de 2010 (MCGLASHAN et al., 2018).

4.2. XHTML+VoiceXML

O XHTML+VoiceXML, também conhecido com X+V, é uma linguagem proposta pelo W3C para o desenvolvimento de aplicações com interface multimodal (visual e voz). O X+V conta com um subconjunto de elementos XHTML, uma linguagem de marcação para criar aplicações visuais, e o VoiceXML, uma linguagem de marcação para criar aplicações com interação por voz (WATANABE et al., 2008).

As linguagens XHTML e VoiceXML são integradas por meio do *Document Object Model* (DOM), que é uma API de programação para HTML e XML proposta pelo W3C. Com o DOM, os programadores podem criar e construir documentos, navegar em sua estrutura e adicionar, modificar ou excluir elementos e conteúdo.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML +Voice//EN" "xhtml+voice10.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xmlns:ev="http://www.w3.org/2001/xml-events"
xmlns:vxml="http://www.w3.org/2001/voicexml20">
   <head>
      <title>Skeleton XHTML+Voice Document</title>
      <!-- voice handlers -->
      <vxml:form id="sayHello">
         <vxml:block>Hello World</vxml:block>
      </vxml:form>
   </head>
   <body>
      <h1>Skeleton XHTML+Voice Document</h1>
      <p ev:event="onclick" ev:handler="#sayHello">
         This is a sample document that illustrates the markup
      structure of a conformant XHTML+Voice document.
      Notice that the default XML namespace is XHTML --and
      consequently, standard HTML element names do not need
      a namespace prefix.  We can add voice-interaction
      specific elements from the Voice XML 2.0 namespace
      using prefix
         <code>vxml</code>
         .  We can attach event
      handlers using prefix
         <code>ev</code>
         .  Clicking
      anywhere on this paragraph results in a welcome
      message being spoken on account of attaching a
         <code>vxml:form</code>
         handler to this paragraph.
      </p>
   </body>
</html>
```

Figura 5. Estrutura básica de um documento XHTML+VoiceXML(AXELSSON et al., 2018).

Como podemos observar na Figura 5, no cabeçalho inicialmente são definidas a versão do XML, DOCTYPE e *Document Type Definition* (DTD). Em seguida temos o elemento <html>, onde estão contidos os elementos <head>, composto pelas funções de voz e o título da página, e <body>, composto por todas informações gráficas HTML. Ainda dentro do elemento <head>, é possível notar o elemento <vxml:form>, que define informações referentes a funções de voz.

4.3. GestureML

O GestureML (GML), também conhecido como Gesture Markup Language, é considerada a primeira linguagem de marcação para interações baseadas em gestos multitoque. A linguagem criada por Gestureworks é baseada em XML, sendo também uma linguagem declarativa (GESTUREML, 2018).

Por se tratar de uma linguagem declarativa, possui alto nível, flexível e extensível, permitindo que os desenvolvedores explorem novos paradigmas de interações e experiências com o usuário (ORTEGA et al., 2013).

Com GML os desenvolvedores projetam livremente interações de gestos multitoque com alta precisão. O gesto capturado pode ser simples, com um único toque, ou complexo, com vários toques em sequências aleatórias. As sequências de gestos podem ser utilizadas para detecção de senha, por exemplo.

A Gestureworks desenvolveu um conjunto gestos descritos em XML conhecido como Gestureworks Core. Este conjunto é uma biblioteca de gestos aberta que inclui definições para dezenas de gestos pré-construídos e possibilidades ilimitadas para novos gestos e sequências de gestos personalizados, e está disponível para as principais

linguagens de programação como C++, C#, Java e Python. A utilização do Gestureworks Core fornece aos desenvolvedores a utilização de uma grande quantidade de gestos, permitindo a reformulação e refinamento de acordo com as necessidades, sendo possível a compilação e distribuição.

Entre as principais vantagens do GML tem-se a facilidade de leitura e estruturação utilizando XML, sendo um modelo unificado para entender e descrever a análise de gestos. A linguagem permite mapeamento de transformação nativa, métodos simples de sequenciamento de gestos, e recurso em tempo real (GESTUREML, 2018).

4.4. SMUIML

Com o objetivo de oferecer uma linguagem que permita descrever interações multimodais de maneira fácil, a *Synchronized Multimodal User Interaction Modeling Language* (SMUIML) foi criada. Com ela, é possível fazer a modelagem de diálogo homem-computador e os diversos eventos associados a este diálogo, e de que maneira estes diálogos serão sincronizados (DUMAS; LALANNE; INGOLD, 2010).

A linguagem SMUIML inicialmente foi projetada como uma linguagem XML, descrevendo cada interação entre o homem e o computador em três diferentes níveis: nível de diálogo, nível de eventos e nível de entrada/saída. Na Figura 6, é possível verificar os diferentes níveis.

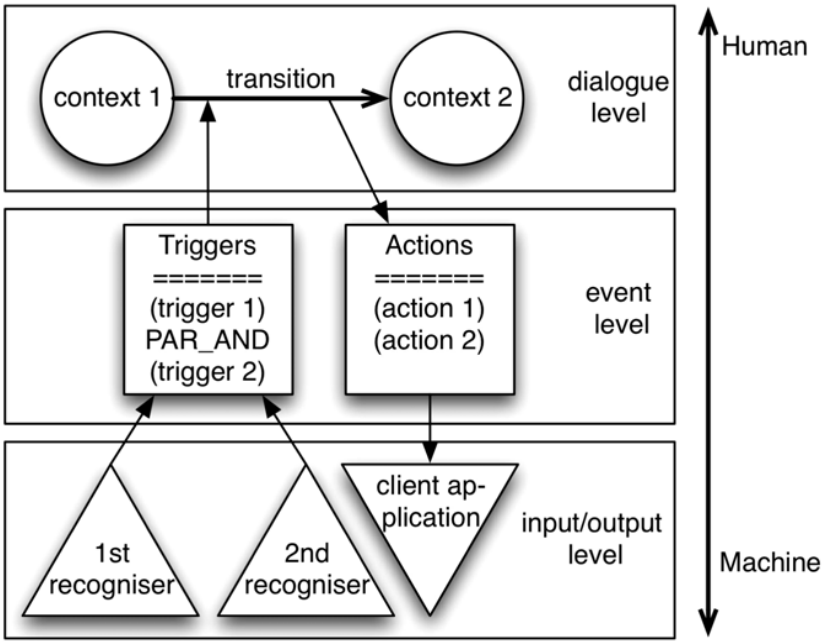

Figura 6. As três camadas de abstração de SMUIML (DUMAS; SIGNER; LALANNE, 2014).

No nível de entrada/saída o desenvolvedor usando a linguagem SMUIML descreve as variáveis de entrada, diferentes modalidades e reconhecedores. Logo acima, no nível de eventos é descrito os acionadores de eventos de entrada (*triggers*) e as ações de saída. O nível mais alto, nível de diálogo, é onde ocorre a descrição dos diálogos entre o homem e a computador por meio de uma máquina de estados finitos.

4.5. NCL

Criada no Laboratório TeleMídia da PUC-Rio, a *Nested Context Language* (NCL) é uma linguagem declarativa utilizada no desenvolvimento de documentos hipermídia baseados no modelo conteitual *Nested Context Model* (NCM) e em XML (AZEVEDO et al., 2011; SOARES et al., 1999; SOARES; RODRIGUES; MUCHALUAT SAADE, 2000). A primeira versão da NCL foi especificada através de um XML DTD (ANTONACCI, 2000). Na versão 2.0, a NCL passou a ser especificado através de um XML Schema, sendo projetada de forma modular, permitindo a combinação de seus módulos em diferentes perfis de linguagem. Estes perfis podem agrupar um subconjunto de módulos NCL, ou ser combinado com módulos de outras linguagens. Também é possível que outras linguagens incorporem NCL em suas linguagens (SAADE, 2003).

A versão atual, NCL 3.0, introduziu a navegação através do uso de teclas e funcionalidades de animação. Além destas duas novas funcionalidades, foram efetuadas modificações na funcionalidade do template de nó de composição, e a reestruturação de conectores hipermídia com o objetivo de permitir uma notação mais concisa (LOPES, 2013).

A NCL permite a definição de metadados, animações e transições de forma simplificada, por ser uma linugagem declarativa. Ao contrário da linguagem SMIL, a NCL permite a reutilização. Outro benefício é a possibilidade de integração com a linguagem Lua.

NCL em principio, não foi desenvolvida para prover suporte a interações multimodais, até porque assim como SMIL e HTML, essas linguagens possuem foco na sincronização de modalidades audiovisuais como texto, gráficos e vídeos e como controle mouse e teclado, o que não inviabiliza extende-las através de outras tecnologias para dar suporte a experiências multimodais assim como realizado por Guedes (GUEDES; AZEVEDO; BARBOSA, 2017).

5. Conclusão

A interação multimodal, modalidade da área de IHC, está presente em vários dispositivos e ferramentas utilizadas pelas pessoas. Para que sejam possíveis estas interações, são necessários alguns métodos de entrada e saída, e os sentidos humanos estão diretamente ligado a estes métodos, envolvendo muitas vezes mais de um deles.

Existem alguns dispositivos intermediários que auxiliam na interação humano computador, como mouse, teclados, controles, câmeras, mas nem sempre há necessidade de utilização de um dispositivo intermediário. Como exemplo da ausência da necessidade de utilização de um dispositivo intermediário, citam-se as telas sensíveis ao toque, onde o usuário interage por meio do sentido do tato com uma tela ou projeção. Nos dois casos, há necessidade de uma aplicação que faz o papel de *middleware* entre o dispositivo final e o usuário.

Nos dispositivos finais, uma infinidade de tipos de mídias são apresentados. Quando há a integração de dois ou mais tipos de mídia, permitindo ao usuário a navegação e utilização necessária, damos o nome de hipermídia.

Toda a integração entre a interação multimodal em aplicações hipermídia é descrita por alguma linguagem. As principais linguagens utilizadas são linguagens

declarativas, que permitem o desenvolvimento de maneira simples e fácil entendimento. Várias linguagens se basearam na linguagem XML, como VoiceXML ou GestureML por exemplo.

Com a constante evolução da tecnologia, espera-se que novos dispositivos desenvolvidos permitam ainda mais a interação multimodal. Junto com a evolução dos dispositivos, as linguagens deverão ser aperfeiçoadas, tanto na facilitação de uma experiência como novas técnicas de programação.

Referências

ALVES, Melissa L. M. et al. Nintendo Wii™ Versus Xbox Kinect™ for Assisting People With Parkinson's Disease. **Perceptual and Motor Skills**, [s. l.], p. 003151251876920, 2018. Disponível em: <http://journals.sagepub.com/doi/10.1177/0031512518769204>. Acesso em: 15 out. 2018.

ANTONACCI, Meire Juliana. **NCL: Uma Linguagem Declarativa para Especificação de Documentos Hipermídia com Sincronização Temporal e Espacial**. 2000. Rio de Janeiro, Brasil, 2000.

AXELSSON, Jonny et al. XHTML+Voice Profile 1.0. . 2018.

AZEVEDO, Roberto Gerson Albuquerque et al. Textual authoring of interactive digital TV applications. In: PROCEDDINGS OF THE 9TH INTERNATIONAL INTERACTIVE CONFERENCE ON INTERACTIVE TELEVISION - EUROITV '11 2011, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 2011. Disponível em: <http://portal.acm.org/citation.cfm?doid=2000119.2000169>. Acesso em: 10 nov. 2018.

BOLT, Richard A. et al. "Put-that-there": Voice and Gesture at the Graphics Interface. In: PROCEEDINGS OF THE 7TH ANNUAL CONFERENCE ON COMPUTER GRAPHICS AND INTERACTIVE TECHNIQUES - SIGGRAPH '80 1980, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 1980. Disponível em: <http://portal.acm.org/citation.cfm?doid=800250.807503>. Acesso em: 15 out. 2018.

CHURCHILL, Elizabeth. The Past, Present and Future of Human Computer Interaction. [s. l.], 2018. Disponível em: <https://repository.kaust.edu.sa/handle/10754/627101>. Acesso em: 15 out. 2018.

DUMAS, Bruno; LALANNE, Denis; INGOLD, Rolf. Description languages for multimodal interaction: A set of guidelines and its illustration with SMUIML. **Journal on Multimodal User Interfaces**, [s. l.], v. 3, n. 3, p. 237–247, 2010.

DUMAS, Bruno; LALANNE, Denis; OVIATT, Sharon. Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In: **Human Machine Interaction**. [s.l.] : Springer-Verlag, 2009. p. 3–26.

DUMAS, Bruno; SIGNER, Beat; LALANNE, Denis. A graphical editor for the SMUIML multimodal user interaction description language. **Science of Computer Programming**, [s. l.], v. 86, p. 30–42, 2014. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167642313001019>

GESTUREML. **GestureML**. 2018. Disponível em: <http://www.gestureml.org/doku.php>. Acesso em: 5 dez. 2018.

GOMES SOARES, Luiz Fernando; MORENO, Marcio Ferreira; GUEDES VASCONCELOS, Alan Lívio. Controlling the focus and input events in multimedia applications. In: PROCEEDINGS OF THE 30TH ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING - SAC '15 2015a, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2695664.2695885>. Acesso em: 12 nov. 2018.

GOMES SOARES, Luiz Fernando; MORENO, Marcio Ferreira; GUEDES VASCONCELOS, Alan Lívio. MultiSEM: A Mulsemedia Model for Supporting the Development of Authoring Tools. In: PROCEEDINGS OF THE 30TH ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING - SAC '15 2015b, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2695664.2695885>. Acesso em: 28 out. 2018.

GUEDES, Álan Lívio Vasconcelos; AZEVEDO, Roberto Gerson de Albuquerque; BARBOSA, Simone Diniz Junqueira. **Extending multimedia languages to support multimodal user interactions**. 2017. PUC-Rio, [s. l.], 2017. Disponível em: <http://www.telemidia.puc-rio.br/publication/2017_09_guedes.html>. Acesso em: 27 maio. 2018.

LAZAR, Jonathan; FENG, Jinjuan Heidi; HOCHHEISER, Harry. **Research methods in human-computer interaction**. [s.l: s.n.]. Disponível em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=hbkxDQAAQBAJ&oi=fnd&pg=PP1&dq=computer+interaction&ots=Sp4889_86W&sig=_noEobX_kEzWUrOws8MJPQd_wiE&redir_esc=y#v=onepage&q=computer interaction&f=false>. Acesso em: 15 out. 2018.

LEE, Wei-Po; KAOLI, Che; HUANG, Jhih-Yuan. A smart TV system with body-gesture control, tag-based rating and context-aware recommendation. **Knowledge-Based Systems**, [s. l.], v. 56, p. 167–178, 2014. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705113003572>. Acesso em: 12 nov. 2018.

LOPES, Diogo Cesar Souza. **WEB COMPOSER: uma proposta de ferramenta web para autoria e execução de aplicações de TV Digital**. 2013. Federal University of Juiz de Fora, [s. l.], 2013.

MAKLEYSTON GOMES PEREIRA, Danne; JOSÉ DA SILVA SILVA, Francisco; LÍVIO VASCONCELOS GUEDES, Álan. **A Middleware Perspective for Integrating Ginga-NCL Applications with the Internet of Things**. [s.l: s.n.]. Disponível em: <http://andsel.github.io/moquette/>. Acesso em: 28 out. 2018.

MATTOS, Douglas Paulo De; MUCHALUAT SAADE, Débora Christina. STEVE. In: PROCEEDINGS OF THE 22ND BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB - WEBMEDIA '16 2016, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 2016. Disponível em: <http://dl.acm.org/citation.cfm?doid=2976796.2976865>. Acesso em: 27 out. 2018.

MATTOS, Douglas P.; MUCHALUAT-SAADE, Débora C. MultiSEM. In:

PROCEEDINGS OF THE 24TH BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB - WEBMEDIA '18 2018, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 2018. Disponível em: <http://dl.acm.org/citation.cfm?doid=3243082.3243114>. Acesso em: 27 out. 2018.

MCGLASHAN, Scott et al. Voice Extensible Markup Language (VoiceXML) 3.0. . 2018.

ORTEGA, Francisco R. et al. Exploring modeling language for multi-touch systems using petri nets. In: PROCEEDINGS OF THE 2013 ACM INTERNATIONAL CONFERENCE ON INTERACTIVE TABLETOPS AND SURFACES - ITS '13 2013, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 2013. Disponível em: <http://dl.acm.org/citation.cfm?doid=2512349.2512400>. Acesso em: 5 dez. 2018.

OSHRY, Matt et al. Voice Extensible Markup Language (VoiceXML) 2.1. . 2018.

RAMADAN, Rabie A.; VASILAKOS, Athanasios V. Brain computer interface: control signals review. **Neurocomputing**, [s. l.], v. 223, p. 26–44, 2017. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231216312152>. Acesso em: 28 out. 2018.

REEVES, Leah M. et al. Guidelines for multimodal user interface design. **Communications of the ACM**, [s. l.], v. 47, n. 1, p. 57, 2004. Disponível em: <http://portal.acm.org/citation.cfm?doid=962081.962106>. Acesso em: 12 nov. 2018.

ROGERS, Yvonne.; SHARP, Helen.; PREECE, Jennifer. **Design de interação : além da interação humano-computador**. [s.l.] : Bookman, 2013.

SAADE, Débora Christina Muchaluat. **Relações em Linguagens de Autoria Hipermídia: Aumentando Reuso e Expressividade**. 2003. Rio de Janeiro, Brasil, 2003.

SCHLÖMER, Thomas et al. Gesture recognition with a Wii controller. **Proceedings of the 2nd international conference on Tangible and embedded interaction TEI 08**, New York, New York, USA, p. 11, 2008. Disponível em: <http://portal.acm.org/citation.cfm?doid=1347390.1347395>. Acesso em: 9 jun. 2016.

SOARES, Luiz Fernando G. et al. Versioning Support in the HyperProp System. **Multimedia Tools and Applications**, [s. l.], v. 8, n. 3, p. 325–339, 1999. Disponível em: <http://link.springer.com/10.1023/A:1009670209489>. Acesso em: 27 out. 2018.

SOARES, Luiz Fernando G.; RODRIGUES, Rogério F.; MUCHALUAT SAADE, Débora C. Modeling, authoring and formatting hypermedia documents in the HyperProp system. **Multimedia Systems**, [s. l.], v. 8, n. 2, p. 118–134, 2000. Disponível em: <http://link.springer.com/10.1007/s005300050155>. Acesso em: 27 out. 2018.

SOUSA, Fernando Henrique. Uma revisão bibliográfica sobre a utilização do Nintendo® Wii como instrumento terapêutico e seus fatores de risco. **Revista Espaço Acadêmico**, [s. l.], v. 8, n. 123, p. 155–160, 2011.

TALARICO NETO, Americo. **Uma abordagem para projeto de aplicações com interação multimodal da Web**. 2011. Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo, São Carlos, 2011. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-07062011-091441/>. Acesso

em: 14 out. 2018.

TURK, Matthew. Multimodal interaction: A review. **Pattern Recognition Letters**, [s. l.], v. 36, p. 189–195, 2014. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0167865513002584>. Acesso em: 4 nov. 2018.

W3C. **VoiceXML's History – VoiceXML**. 2018.

WATANABE, Willian Massami et al. Desenvolvimento de componentes de interfaces multimodais ricas para a web utilizando X+V e Dojo widgets. In: COMPANION PROCEEDINGS OF THE XIV BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB - WEBMEDIA '08 2008, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 2008. Disponível em: <http://portal.acm.org/citation.cfm?doid=1809980.1810007>