REVISTA CIENTÍFICA ELETRÔNICA DE SISTEMAS DE INFORMAÇÃO - ISSN 1807-1872
PUBLICAÇÃO CIENTÍFICA DA FACULDADE DE CIÊNCIAS JURÍDICAS E GERENCIAIS DE GARÇA/FAEG
ANO II, NÚMERO, 04, FEVEREIRO DE 2006. PERIODICIDADE: SEMESTRAL

RECONHECIMENTO E SINTETIZAÇÃO DE VOZ USANDO JAVA SPEECH

MARANGONI, Josemar Barone
Docente da Faculdade de Ciências Gerenciais e Jurídicas de Garça – FAEG/Garça
josemarbarone@gmail.com

PRECIPITO, Waldemar Barilli
Docente da Faculdade de Ciências Gerenciais e Jurídicas de Garça – FAEG/Garça
waldemarbp@terra.com.br

RESUMO

Java Speech é uma ferramenta valiosa que pode ser utilizado para auxiliar várias pessoas no trabalho do dia-a-dia. Neste estudo buscou-se verificar o reconhecimento e também o sintetizador da fala, apresentando suas vantagens e desvantagens e, ainda, quando deve ser utilizado ou não.
Palavras-chave: Java Speech; fala.

ABSTRACT

Java Speech is a valuable tool where it can be used to assist some people in the day work. The study of Java was about recognizer and synthesizer and also of it speaks, where they present its advantages and disadvantages, when it must be used or not, its problems and benefits.
Keywords: Java Speech; speaks.

1 Java Speech API

As empresas podem beneficiar-se de uma larga escala de aplicações da tecnologia da fala usando o Java Speech API. Por exemplo, os sistemas interativos da resposta de voz são uma alternativa atrativa às relações sobre o telefone; os sistemas de ditado podem ser

consideravelmente mais rápidos do que a entrada datilografada para muitos usuários; a tecnologia de reconhecimento da fala melhora a acessibilidade aos computadores para muitas pessoas com limitações físicas.

O Java Speech API define um padrão para a utilização da fala para interação com o computador. Duas tecnologias de fala são suportadas pelo Java Speech API. Uma delas é o reconhecimento de fala (reconhecimento da fala) e a outra é o sintetizador de fala (síntese da fala). O reconhecimento de fala fornece aos computadores a habilidade de “escutar” a língua falada e de determinar o que foi dito, ou seja, processa a entrada de áudio que contém a fala convertendo para texto. O sintetizador de fala, por sua vez, fornece o processo reverso de produzir a fala sintética do texto gerado por uma aplicação, por um applet ou por um usuário. É chamada freqüentemente como a tecnologia text-to-speech (texto para fala).

1.1 Aplicações permitidas no Java Speech

As potencialidades da plataforma de Java fazem atrativo para o desenvolvimento de uma larga escala de aplicações. Com a adição do Java Speech API, os programadores das aplicações em Java podem estender e complementar relação de usuário existentes com entrada e saída de comunicação. Para programadores desse tipo de aplicações de fala, a plataforma de Java oferece agora uma alternativa atrativa com:

- Portabilidade: a linguagem de programação Java, os APIs e as máquinas virtuais estão disponíveis para uma larga variedade de plataformas de hardware e de sistemas operacionais e são suportados pela maioria dos navegadores.

- Ambiente poderoso e compacto: a plataforma Java fornece aos programadores uma poderosa orientação a objeto, utilizando “garbage

collection” que permite um desenvolvimento rápido e a confiabilidade melhorada.

- Rede segura: a plataforma de Java inclui uma rede de segurança robusta.

1.2 Java Speech e outros Java APIs

O Java Speech API é um dos Java Media APIs, uma relação de software que fornece acesso a plataforma de áudio, vídeo, gráficos multimídia, gráficos 2D e 3D, animações, telefonia, imagem avançada, e mais. O Java Speech API, em combinação com os outros meios APIs de Java, permite que os programadores enriqueçam aplicações de Java com meios e potencialidades ricos de uma comunicação que se encontram com as expectativas de usuários de hoje, e pode realçar uma comunicação pessoa-a-pessoa.

1.2.1 Aplicações da tecnologia Speech

A tecnologia Speech está tornando-se cada vez mais importante nas empresas que computam enquanto é usada para melhorar relações de usuário existentes e suportar meios novos da interação humana com computadores. A tecnologia do speech permite o uso “hands-free” (mãos livres) dos computadores e suporta o acesso aos computadores fora da mesa de trabalho. O reconhecimento de fala e o sintetizador de fala podem melhorar a acessibilidade do computador para usuários com inabilidade e podem reduzir o risco de ferimento repetitivo da tensão e outros de problemas causados por relações atuais.

1.3 Desktop e sistemas de telefone

A tecnologia speech pode aumentar as relações de usuário com os gráficos, pode ser usada para fornecer alertas audíveis com as respostas faladas de "Sim/Não/OK" que não tiram a atenção do usuário no que ele está fazendo.

Por exemplo, editando um texto no Word o comando "use o tamanho 12, negrito Times new roman" substitui seleções múltiplas do menu e cliques do mouse.

Aplicações em que as mãos ficam ocupadas podem ser realçados usando comandos da fala em combinação com ações do mouse e do teclado e melhorar a velocidade em que os usuários podem manipular objetos. Por exemplo, ao arrastar um objeto, um comando do speech poderia ser usado para mudar seu tipo a cor e de linha sem mover o ponteiro para a barra de menu ou uma paleta da ferramenta.

Os comandos da linguagem natural podem fornecer melhorias na eficiência mas estão sendo usados cada vez mais em ambientes desktop. Para muitos usuários é mais fácil e mais natural produzir comandos falados do que para recordar a posição das funções nos menus e nas caixas de diálogo.

Em muitas situações onde a entrada de teclado é pouco prática e as exposições visuais são restritas, a fala pode fornecer a única maneira interagir com um computador. Por exemplo, os cirurgiões e a outra equipe de funcionários médica podem incorporar relatórios quando suas mãos são ocupadas e quando tocar em um teclado representa um risco da higiene. Em um veículo ou em uma manutenção de linha aérea, armazenando e muitas outras tarefas de "mãos ocupadas", as relações de fala podem fornecer a entrada e a saída prática de dados e podem permitir treinamento por computador.

A tecnologia está sendo usada por muitas empresas para segurar chamadas de cliente e pedidos internos para o acesso à informação, aos recursos e aos serviços.

Por exemplo: "eu tenho e-mail?" "sim, você tem 7 mensagens incluindo 2 mensagens de alta prioridade do gerente de produção." "leia-me por favor o correio do gerente de produção." o "e-mail chegou em 12:30pm..." e assim por diante.

2 Tecnologia Speech

Apesar do investimento muito substancial na pesquisa da tecnologia de reconhecimento de fala nos últimos 40 anos, o sintetizador de fala e as tecnologias do reconhecimento de fala têm ainda limitações significativas. O mais importante, a sintetizador de fala não se encontra sempre com as expectativas elevadas dos usuários familiares com uma comunicação de fala humano à humano-natural. Compreender as limitações é importante para o uso eficaz da entrada e da saída da fala em uma relação de usuário e para compreender algumas das características avançadas do Java Speech API.

2.1 Síntese da fala

Um sintetizador de fala (síntese da fala), converte o texto escrito na língua falada. A síntese da fala é também referenciada como a conversão TTS (text-to-speech).

As principais etapas de se produzir um texto são:

- Análise da estrutura: processa o texto de entrada para determinar onde os parágrafos, as sentenças e outras estruturas começam e terminam. Para a maioria das línguas, os dados da pontuação e do formato são usados neste estágio.

- Pré-processamento do texto: analisa o texto de entrada para construções especiais da língua. Em inglês, tratamentos especiais são requeridos para as abreviaturas, acrônimos, datas, épocas, números, moeda corrente, endereços de e-mail e muitos outros formulários. Outras línguas necessitam processar especial para estes formulários e a maioria das línguas tem outras exigências especializadas.

A próxima etapa é a conversão do texto em fala que é assim:

- Conversão do Texto ao fonema: converte cada palavra aos fonemas. Um fonema é uma unidade básica do som em uma língua. O inglês dos Estados Unidos tem ao redor 45 fonemas incluindo os sons da consoante e da vogal. Diferentes línguas têm conjuntos diferentes de sons (fonemas diferentes). Por exemplo, o japonês tem poucos fonemas incluindo os sons não encontrados em inglês como, por exemplo, o 'ts' de 'tsunami'.

- Análise de Prosody: processe a estrutura de sentença, as palavras e os fonemas para determinar o prosody apropriado para a sentença. Prosody inclui muita das características da fala a exceção dos sons das palavras que estão sendo faladas. Isto inclui o passo (ou a melodia), o sincronismo (ou o ritmo), pausar, a taxa faladora, a ênfase em palavras e muitas outras características.

- Produção do waveform: finalmente, os fonemas e a informação prosody são usados produzir o waveform para cada sentença. Há muitas maneiras em que a fala pode ser produzido, do fonema e informação prosody. A maioria dos sistemas atuais faz em uma de duas maneiras a concatenação dos pedaços da fala humano gravado, ou síntese do formato usando as técnicas processando de sinal baseadas no conhecimento de como o som dos fonemas e de como prosody afeta aqueles fonemas. Os detalhes da geração do waveform não são tipicamente importantes para os programadores desse tipo de aplicação.

2.2 Limitações do Sintetizador de fala (Síntese da fala)

O speech sintetizadores (síntese da fala) pode cometer erros. As orelhas humanas são bem ajustadas para detectar estes erros, assim o trabalho cuidadoso de programadores pode minimizar erros e melhorar a qualidade da saída da fala.

O Java Speech API e o Java Speech Markup Language (JSML) fornece muitas maneiras para um programador de aplicação melhorar a qualidade da saída de um sintetizador de fala. O capítulo 6 descreve técnicas de programação para controlar uma síntese com o Java Speech API. Algumas de suas características que realçam a qualidade incluem:

* Habilidade de marcar o começo e o fim dos parágrafos e das sentenças.

* Habilidade de especificar pronunciações para alguma palavra, acrônimo, abreviatura ou a outra representação especial do texto.

* Controle explícito das pausas, dos limites, da ênfase, do passo, da taxa faladora e do loudness para melhorar a saída prosody.

Estas características permitem que um programador ou um usuário cancele o comportamento de um sintetizador de fala para corrigir a maioria dos erros potenciais. A seguir uma descrição de algumas das fontes de erros:

* Análise da estrutura: a pontuação e o formato não indicam consistentemente onde os parágrafos, as sentenças e outras estruturas começam e terminam. Para o exemplo, o ponto final em "EUA." pode não ser interpretado
como o fim de uma sentença.

* Pré-processamento de texto: não é possível para um synthesizer saber todas as abreviaturas e acrônimos de uma língua. Não é sempre possível para um synthesizer determinar como processar datas e épocas, por exemplo, são "8/5" "oitavo dia de maio" ou "quinto dia de

agosto"? Se "1998" estiverem lidos como "mil novecentos e noventa e oito" (como um ano), como "mil novecentos e noventa e oito" (um número regular) ou como "um nove nove oito" (parte de um número de telefone).

* Conversão de Texto à fonema : a maioria dos synthesizers podem pronunciar dez dos milhares ou mesmo das centenas dos milhares das palavras corretamente. Entretanto, há sempre as palavras novas que deve supor para (nomes especiais apropriados para povos, companhias, produtos, etc.), e as palavras para que a pronunciação está ambígua.

* Análise de Prosody: frasear corretamente uma sentença, para produzir a melodia correta para uma sentença e para enfatizar corretamente palavras, requer idealmente uma compreensão do sentido das línguas que os computadores não possuem. Ao invés dos sintetizadores de falas tentarem supor o que um ser humano pode produzir e, às vezes, as suposições são artificiais e não naturais.

* Produção do waveform: sem bocas, pulmões ou outro instrumento da fala humano, um sintetizador de fala produzirá freqüentemente uma fala que soa artificial, mecânico ou de outra maneira diferente da fala humano. Em algumas circunstâncias um som robótico é desejável, mas para a maioria das falas que das aplicações isso soa como perto do ser humano porque é possível e mais fácil de compreender e mais fácil de escutar por longos períodos de tempo.

2.3 Avaliação do Sintetizador de fala

Os seres humanos são condicionados por uma vida de escutar e do falar. Os ouvidos humanos (e o cérebro) são muito sensíveis às mudanças pequenas na qualidade da fala. Um ouvinte pode detectar as mudanças que puderam indicar o estado emocional de um usuário, um

sotaque, um problema da fala ou muitos outros fatores. A qualidade da síntese da fala atual remanesce abaixo daquela da fala humano, assim os ouvintes devem fazer mais esforço do que o normal para compreender a fala e ignorar os erros. Para usuários novos, escutar um sintetizador de fala por períodos prolongados pode ser cansativo e insatisfatório.

2.4 Reconhecimento de fala

Reconhecimento de fala é o processo de converter a língua falada ao texto escrito ou a algum formulário similar. As características básicas de um identificador da fala que suporta a fala API de Java são:

É mono-lingual: suporta uma única língua especificada.

Processa uma única entrada de áudio.

Pode opcionalmente adaptar-se à voz de seus usuários.

Suas gramáticas podem ser dinamicamente atualizadas.

Tem um conjunto pequeno, definido de propriedades aplicação - controle.

As etapas principais de um identificador típico de fala são:

Projeto da gramática: as gramáticas definem as palavras que podem ser faladas por um usuário e pelos testes padrões em que podem ser faladas. Uma gramática deve ser criada e ativada para que um identificador saiba o que deve aguardar até escutar no áudio de entrada.

Processador de sinal: analisa as características do spectrum (freqüência) do áudio de entrada.

Reconhecimento do fonema: comparam os testes padrões do spectrum aos testes padrões dos fonemas da língua que está sendo reconhecida. (uma breve descrição dos fonemas é fornecida na seção da síntese de fala na discussão da conversão do texto ao fonema).

Reconhecimento de palavras: comparam a seqüência de fonemas prováveis de encontro às palavras e aos testes padrões das palavras especificadas pelas gramáticas ativas.

Geração de resultado: fornece a aplicação com a informação sobre as palavras que o identificador detectou no áudio de entrada. A informação do resultado será fornecida sempre uma vez que o reconhecimento de um única sentença está completo, mas pode também ser fornecida durante o processo do reconhecimento. O resultado indica sempre a melhor suposição para o identificador de que o usuário tenha dito, mas pode também indicar suposições alternativas.

Java Speech API suporta dois tipos básicos de gramática: gramáticas da regras gramaticais e do ditado. Estes tipos da gramática diferem na maneira em que as aplicações ajustam as gramáticas, nos tipos de sentenças que permitem, na maneira em que os resultados são fornecidos, na quantidade de recursos computacionais requeridos, e na maneira em que são usados eficazmente no projeto da aplicação.

3 Conclusões

O trabalho teve como sua maior importância o estudo do Java Speech API. Concluiu-se que há muito que fazer ainda para chegar a ajudar deficientes físicos, crianças ou idosos.

Como projeto futuro poderia ser feito um aprimoramento do speech, para melhor reconhecimento da voz.

4 Referências bibliográficas:

http://java.sun.com/products/java-media/speech/forDevelopers/jsapi-guide/Preface.html