RECONHECIMENTO DE VOZ

ÍNDICE

1 INTRODUÇÃO

O objetivo do presente trabalho é o de demonstrar o conceito de sistemas computacionais de reconhecimento de voz, a sua utilidade dentro do atual estágio da evolução humana e o impacto que causará a sua utilização maciça num futuro próximo. É ponto pacífico que as atuais tecnologias de interface homem/computador para entrada de dados, como teclado e mouse, estão fadadas a ser substituídas por outras, mais naturais aos seres humanos, principalmente a fala, já que é a principal forma de comunicação e, por excelência, a melhor, digamos assim, interface homem/homem que existe.

Desde "2.001: Uma Odisséia no Espaço", o homem sonha com o dia em que poderá falar (e ser respondido) com o computador, estabelecendo assim com ele um contato completamente diferente de tudo o que vimos até hoje nesse campo. O ano 2.000, no entanto, já está aí e ainda falta muito para termos o nosso HALL, tal e qual foi idealizado na ficção científica por Arthur C. Clarke e Stanley Kubrick. Ainda assim,

muito já se avançou nessa área e é sobre isso que se irá discorrer nas próximas páginas deste trabalho.

Retornar

---

2 RECONHECIMENTO DE VOZ - O QUE É?

A linguagem falada é a forma mais usada de comunicação entre os seres humanos. Devido à capacidade do nosso cérebro de interpretar informações extremamente complexas, podemos, de forma praticamente inconsciente, captar facilmente em uma mensagem falada várias informações além da transmitida textualmente pelas frases vocalizadas. Reconhecemos assim quem nos está falando (o agente emissor da mensagem), sua posição no espaço físico, seu estado emocional e vários outros dados que podem estar escondidos no tom de voz usado (ironia, seriedade ou tristeza, por exemplo).

As máquinas que nos rodeiam, e o computador dentre elas, no entanto, não tem condições de analisar esse tipo de dados como nós o fazemos. Nossa relação com as máquinas, então, é relativamente "frio". Assim, para nos comunicarmos com as máquinas precisamos apertar botões, mover alavancas, digitar comandos em teclados, em suma, formas bastante estranhas ao nosso meio normal de comunicação, a fala. Por isso, para podermos trabalhar com as máquinas (e isso é essencial no atual estágio de desenvolvimento da nossa civilização) devemos nos adaptar à elas e aprender uma série de métodos diferentes da nossa natureza humana e, portanto, incômodos, principalmente para o cidadão médio e pouco afeito às evoluções tecnológicas.

Para melhorar a relação homem/máquina, ou a interface entre eles, os cientistas têm pesquisado há tempos tecnologias que permitissem o reconhecimento de voz, habilitando assim qualquer pessoa a comunicar-se diretamente e de forma natural com os computadores e os sistemas computacionais que controlariam as máquinas que nos cercam no dia-a-dia. Isso representaria um avanço enorme por vários motivos: a menor curva de aprendizado necessário para a utilização de um sistema ou equipamento, a rapidez e eficiência com que essa utilização se daria, o aumento do desempenho individual e da produção e a possibilidade de estar com as mãos livres para se fazer outras coisas enquanto se utiliza o computador ou o equipamento dotado dessa tecnologia. Até mesmo a segurança pode ser implementada com o desenvolvimento do reconhecimento de voz, já que possibilita a identificação pessoal de indivíduos através do timbre de sua voz, permitindo ou negando a sua entrada em uma área reservada de uma empresa, por exemplo.

A tecnologia de reconhecimento de voz vem sendo pesquisada por várias empresas e instituições para ser viabilizada em vários campos, tanto na área comercial quanto na

estatal. O governo americano, por exemplo, vem trabalhando em um sistema de reconhecimento de voz para a realização do censo por telefone. O cidadão seria atendido por um computador, responderia verbalmente os dados solicitados e o sistema extrairia automaticamente as informações necessárias das frases e preencheria com elas o formulário do censo, economizando ao Estado americano uma quantia de 2 a 4 bilhões de dólares. A empresa americana IBM vem implementando há alguns anos um software utilitário presente em seu sistema operacional OS/2, chamado VoiceType, que realiza reconhecimento de voz numa tentativa de permitir a emissão de comandos para o computador e outras coisas interessantes como, por exemplo, o ditado de documentos diretamente para o processador de textos, sem a necessidade de digitação. Os resultados vêm sendo altamente promissores, apesar da ainda alta taxa de ineficiência apresentada.

Assim, observa-se grandes avanços ocorrendo a todo momento nessa área, permitindo a previsão de ótimos resultados num futuro não tão remoto, alavancando setores da tecnologia ainda não explorados e propiciando um novo viés para os estudos e pesquisas que se farão mais adiante.

Retornar

---

3 A TECNOLOGIA E O RECONHECIMENTO DE VOZ

O reconhecimento de voz é o processo no qual se extrai de forma automática a informação necessária do sinal de voz. Essa informação, como já vimos, pode ser uma dentre várias as que se apresentam codificadas dentro do sinal de voz emitido. O sistema pode estar interessado na informação textual da fala para a edição de um texto ou execução de um comando. O sistema também pode estar interessado em algumas palavras chave, como "sim" ou "não", que possam estar inseridas na frase. Ou ainda no timbre pessoal do agente emissor do comando de voz, para a identificação do mesmo como uma medida de segurança.

O processo de reconhecimento de voz pelo sistema computacional ocorre em três fases distintas: aquisição do sinal de voz, extração de parâmetros e reconhecimento do padrão.

A aquisição do sinal de voz se dá à partir de um dispositivo conversor analógico/digital, que obtém o sinal a ser reconhecido. Em um microcomputador de mesa, esse processo poderia ser o seguinte: um usuário emitiria o sinal de voz em um microfone acoplado à uma placa de som que digitaliza o sinal analógico deixando-o preparado para a próxima fase do processamento.

Na segunda fase, um algoritmo de parametrização, ou seja, um programa desenvolvido para tratar de forma parametrizada o dado de entrada (comando vocal) através de um conjunto de características que descrevem de maneira adequada as propriedades do sinal da voz, extraindo e representando os parâmetros necessários para a sua utilização.

Após a extração das características do padrão, o reconhecimento do padrão responsabiliza-se pela identificação dos mesmos, isto é, através de comparações sucessivas, ele verifica a que padrão de referência (conhecido) o padrão de entrada (o qual se deseja reconhecer) se assemelha.

Nesse ponto se encontram os maiores problemas ainda não solucionados no campo do reconhecimento de voz. Isso porque não se trata somente da simples comparação de dados para se identificar o padrão, já que ele não é fixo. Ocorrem variações nas características da fala que dificultam extremamente o processamento. Suponhamos um usuário emitindo um comando de voz ao computador. Cada vez que ele fizer isso, ocorrer´ pequenas diferenças nas formas de onda que compõem o comando, devido à articulação dos órgãos do aparelho fonador. O reconhecimento do comando também pode ser prejudicado pela distância entre o usuário emissor do sinal de voz e do dispositivo de recepção do mesmo ou por algum obstáculo que se interponha entre os dois. O sinal de voz também pode reverberar na parede, ou possuir variações de amplitude (volume menor ou maior),ou ainda ser atrapalhado pela emissão de outro sinal de voz no mesmo recinto por outra pessoa, entre muitíssimos outros problemas ocorrer . A pronúncia do usuário também pode não ser suficientemente correta devido a um sotaque exótico ou algum problema de dicção.

Além de todos esses problemas, ainda há o da segmentação da fala, o qual tem tomado muito tempo dos cientistas que pesquisam o reconhecimento de voz. Não se tem precisamente uma forma de limitação dos fonemas (menor unidade da fala), o que dificulta o reconhecimento de fala contínua. Porque quando falamos, o som que emitimos é algo como: "Abraoprogramadediçãodetexto" e não "Abra-o-programa-de-edição-de-texto".

Nós, seres humanos, conseguimos facilmente reconhecer todos esses problemas e apreender a informação desejada facilmente e de forma automática de vozes dos mais diversos timbres, amplitudes, com os sotaques mais estranhos e no meio de um emaranhado de ruídos (até certo ponto, é claro). Porém não sabemos exatamente que mecanismos o nosso cérebro utiliza para que esse processamento seja realizado e, portanto, não sabemos como o processador deve se comportar para obter o mesmo sucesso.

As restrições acima irão influenciar características como precisão, tipo de aplicação, custo, entre outras. Para contornar algumas restrições foram determinados certos fatores para o reconhecimento:

- Dependência do Locutor: se o sistema somente reconhece a voz dos locutores

para que foi treinado, tal sistema é dependente do locutor. O sistema é independente do locutor quando é capaz de reconhecer qualquer locutor que não tenha sido treinado;

- Tipo de fala: pode-se reconhecer palavras isoladas ou fala contínua. No primeiro caso é necessário um período mínimo de silêncio entre as palavras pronunciadas e no segundo esta restrição não é aplicada;
- Tamanho do vocabulário: o tamanho do vocabulário influencia a precisão do sistema de reconhecimento. Isto ocorre devido a possível ambigüidade das palavras (palavras semelhantes para o algoritmo classificador).

Retornar

---

4 HARDWARE ESPECÍFICO NO RECONHECIMENTO DE VOZ

Uma das áreas atuais mais promissoras do mercado de informática é o reconhecimento de padrõs vocais, devido a sua grande utilidade em vários segmentos da indústria ou comércio, onde a entrada ou saída de informações beneficiem o trabalho tanto na velocidade de execução de um comando ou de entrada de dados. Nesse contexto, o avanço da tecnologia VLSI têm sido significativo, o que vem permitir a construção de placas processadoras de sinais do tipo DSP com qualidade compatível com as necessidades de software de reconhecimento de padrões cada vez mais rápidos e precisos.

Com o processamento do sinal sendo realizado por hardware específico, o trabalho de reconhecimento se limitaria mais ao problema de comparação dos padrões. Em vista disso, os trabalhos nesta área se orientam por esta abordagem, devido a dificuldade de implementação de certas aplicações de reconhecimento de voz.

As pesquisas no reconhecimento de voz com vocabulário restrito, independente do locutor, e reconhecimento de locutor independente do texto têm abrangido vários tipos de abordagens. Na fase de extração de características, os algoritmos mais usados são o LPC (Linear Predictive Coding - Codificação Preditiva Linear), o modelo mistura Gaussiano) e o FFT (Fast Fourier Transform - Transformada Rápida de Fourier).

Os trabalhos que utilizam LPC consideram a premissa que o trato vocal humano não é fixo. Isto é, de acordo com os tipos de sons emitidos (vogais, consoantes), o aparelho vocal realiza um trato diferente para cada som. Além disto, o trato vocal é particular a cada pessoa, e por isso pode ser um bom parâmetro para reconhecimento de locutor. De encontro a isto, o algoritmo LPC consegue aproximar o seu processamento para modelar o trato vocal. Isto é resultado de seu processamento baseado na análise das

entradas anteriores e um possível erro para predizer o próximo sinal.

O modelo de mistura Gaussiano é usado no reconhecimento de locutores devido à duas grandes interpretações: em primeiro lugar, os componentes individuais do modelo Gaussiano podem ser usados para representar algumas classes acústicas amplas. Estas classes acústicas refletem o trato vocal, os quais são úteis para modelar um locutor. Em segundo, a densidade da mistura Gaussiana é empregada para prover uma suave aproximação para a distribuição das amostras de termo-longo subjacentes das observações obtidas das articulações de um dado locutor.

No reconhecimento de palavras isoladas o algoritmo FFT é muito usado devido a modelagem que realiza do sinal. Tal algoritmo realiza a transferência de abordagem do sinal em função do tempo para sinal em função das freqüências, que pode ser um bom parâmetro para realizar reconhecimentos independentes do locutor.

Na fase de classificação, o problema a ser encontrado pela maioria dos reconhecedores é o tempo de execução, isto é, a velocidade em que o algoritmo de classificação realiza o processamento e chegar a uma solução. Nos métodos matemáticos convencionais a necessidade de processamento é muito grande. Em vista disso, existe uma grande tendência de uso de algoritmos "inteligentes". Tais métodos baseiam-se na técnica de aprendizagem e processamento humano, devido a sua grande capacidade de processamento de informações. Destes métodos pode-se destacar as Redes Neurais Artificiais.

O campo de reconhecimento de fala apresentou um desenvolvimento considerável nos últimos anos, através da utilização de redes neurais e sistemas híbridos, como por exemplo o uso de Redes Neurais e o Modelo de Hidden Markov, que incrementou o nível de reconhecimento para valores acima de 95%, tratando-se do problema de reconhecimento de fala contínuo e independente de interlocutor.

A otimizaçã global torna o sistema mais flexível incorporando várias técnicas propostas anteriormente com o objetivo de apresentar uma menor quantidade de atribuições falsas. O sistema proposto por Bengio et al apresenta um sistema híbrido em que os frames de fala são produzidos por uma combinação de análise do sinal e utilizando redes neurais, sendo que os frames de fala servem como entrada de um sistema que utiliza o Modelo de Hidden Markov. A rede neural é treinada para utilizar incrementalmente os frames de fala, utilizando o algoritmo de Backpropagation, fazendo com que o sistema treine o Modelo de Hidden Markov e a rede neural simultaneamente. Os sistemas independentes de interlocutor são os mais desejados porque tornam possível sua utilização por pessoas que não foram incorporadas no processo de treinamento do sistema.

Os sistemas independentes de interlocutor possuem erros de 2 a 3 vezes superiores aos sistemas dependentes de interlocutor. As tarefas de reconhecimento de fala são divididas em três tipos:

- Reconhecimento isolado de palavras;
- Reconhecimento contínuo de fala;
- Reconhecimento de palavras marcadas.

O reconhecimento de palavras isoladas é o processo mais fácil de reconhecimento de fala. Neste processo somente são reconhecidos palavras isoladas pronunciadas por um interlocutor. Este processo é o mais fácil de ser realizado na medida que o sistema não necessita detectar o final da pronúncia de uma palavra, tarefa complexa em muitos casos.

O processo de palavras marcadas é utilizado em aplicações em que somente uma pequena quantidade de palavras deve ser efetivamente reconhecida dentro de uma sentença pronunciada pelo interlocutor (por exemplo: as palavras "sim" e "não").

Retornar

---

5 REDES NEURAIS

No cérebro humano, centenas ou milhares de neurônios processam informações sobre um mesmo assunto ao mesmo tempo e enviam os resultados para os neurônios seguintes nas sinapses, gerando assim uma análise mais completa e detalhada sobre o assunto examinado do que cada neurônio teria condições de executar isoladamente.

O neurônio artificial, ou nó neural, é um processador simples. Ele obtém as informações do exterior ou de outros nós, toma uma única decisão e, por meio de um único canal de saída, passa o resultado para o nó seguinte. Quando vários nós são ligados em rede, o efeito conjunto é a capacidade de tomar decisões complexas.

A rede neural não é uma configuração especial de hardware e software. É uma abstração matemática que pode ser implementada em micros ou computadores especializados. A maior parte do trabalho de redes neurais se faz com equipamentos especializados pois a velocidade de processamento é de uma extrema importância devido à complexidade dos cálculos.

Os computadores que usam redes neurais executam novas tarefas sem precisar ser reprogramados; eles mesmos fazem esse papel. Os programadores apenas estabelecem objetivos e corrigem o computador até ele ser capaz de resolver o problema sozinho.

Apesar das dificuldades no treinamento de redes neurais e do fato de elas poderem

aprender conceitos errados, esta tecnologia tem a grande vantagem de aprender com as novas situações que se deparam podendo, após um período de treinamento, dar seguimento sem auxílio humano. Isso lhe dá condições de controlar certos programas especiais como esforços de exploração muito longos, extremamente curtos ou muito perigosos para os pesquisadores humanos.

Isso torna as redes neurais a ferramenta ideal para a implementação de sistemas de reconhecimento de voz já que a possibilidade de aprendizado e de comportamento inteligente indicam caminhos a seguir na tentativa de superar os obstáculos que se apresentam no percurso da evolução dessa tecnologia. Dessa forma, sempre que se apresentar uma nova circunstância não prevista pelo sistema, este pode solicitar informações ao usuário, analisar a situaçatilde;o e aprender com ela, registrando o resultado para futuras intervenções.

Retornar

---

6 CONCLUSÃO

Após a pesquisa que deu origem ao presente trabalho, conclui-se que as tecnologias de reconhecimento de voz, tanto na área de software quanto de hardware, vem sendo desenvolvidas a passos largos, criando uma grande expectativa quanto à sua futura utilização e quanto ao impacto que causarão em todos: usuários, empresas e instituições de ensino. O mundo realmente não será o mesmo.

O principal, porém, é que tudo isso criará inúmeras portas para o surgimento de muito mais revoluções, tanto na forma em que as indústrias de tecnologia trabalham, quanto na nossa relação enquanto usuários e os computadores. Nosso relacionamento com eles seré outro completamente diferente, mais direto, mais natural e mais produtivo.

Pessoas com necessidades físicas especiais, como os deficientes visuais e os tetraplégicos poderão se utilizar desses sistemas para trabalharem tranqüilamente com computadores, sem necessidade da utilização dos atuais meios de entrada e saída de dados. Conversando com o computador, eles serão capazes de fazer o trabalho de praticamente qualquer pessoa de forma tão produtiva quanto.

Isso sem contar com o fato de que os computadores nãto serão as únicas máquinas a utilizar os sistemas de reconhecimento de voz. Carros guiados pela fala, elevadores que movem-se para um andar específico apenas com uma palavra e luzes que acendem ou apagam ao som da voz são apenas alguns dos exemplos do que está por vir pela frente para facilitar a vida do ser humano.

Em suma, o futuro que se descortina traz inúmeras novidades mas uma boa parte delas passa necessariamente pelo reconhecimento de voz.

Retornar

---

7 REFERÊNCIAS BIBLIOGRÁFICAS

- Free Speech Journal - Phoneme Probality Estimation with Dynamic Sparsely Connected Artificial Neural Network - Nikos Drakos - http://cslu.cse.ogi.edu/fsj/html/.
- REVOX - UFRGS, UCS, Elevadores SÛR, Fundação Centro Tecnológico para Informática - http://www.ucs.tche.br/revox/.

Retornar