CAPÍTULO 2 - FUNDAMENTAÇÃO TEÓRICA

2.1 - A INTELIGÊNCIA ARTIFICIAL E O RECONHECIMENTO DE FALA

Conforme KAUTZ (Apud COELHO, 1990), o reconhecimento de fala não é exatamente um problema fácil de ser resolvido. KAUTZ incluiu o reconhecimento de fala juntamente com outras dezenas de itens em uma lista que considerou problemas difíceis de serem resolvidos e, dessa forma, ficou atribuído como meta da inteligência artificial o reconhecimento de fala. Diversos autores de livros e revistas têm defendido a hipótese de que computadores já reconhecem a fala da mesma forma que uma padaria anuncia pão para vender. Entretanto, a maioria dos sistemas reconhecedores de fala não passam de um mero *souvenir*, que muitas vezes acaba vindo como brinde na compra de um computador ou algum software específico.

A exemplo disto, o software *brinquedo-brinde* mais famoso é o Voice Pilot da empresa americana Microsoft. Esse reconhecedor pode lidar com até 256 palavras diferentes que, basicamente, acionam comandos do ambiente Windows. O Voice Pilot emprega um tipo discreto de reconhecedor de fala que se apóia em fonemas (QUAIN, 1994). Ainda de acordo com QUAIN (1994), o emprego desta técnica exige menor capacidade de computação, porém a precisão do reconhecedor é sacrificada.

Essa dificuldade de reconhecer a fala associada com o futuro promissor que o reconhecimento de fala poderá proporcionar à humanidade, fez com que 5,6% dos investimentos em pesquisas direcionadas em Inteligência Artificial atingisse o estudo da linguagem natural em 1990 (COELHO, 1990). Esse valor foi levantado conforme uma pesquisa realizada em 21 países, inclusive o Brasil. Essa subdivisão da inteligência artificial, o estudo da línguagem natural, absorve, ainda, a semântica, a interpretação, a geração e a aquisição da linguagem natural.

Quanto aos investimentos, em virtude das freqüentes ramificações que ocorrem dentro desta ciência, é possível que, atualmente, esse valor de 5,6 % esteja menor. Em 1981, na corrida da compreensão da linguagem natural, esse investimento era de 16,2 %, perdendo apenas para os investimentos ocorridos no campo da visão, que na época eram de 17,8 %. A partir de 1987, a visão passou a fazer parte da robótica, que, dessa forma, absorveu todo o seu investimento. Até 1990, os maiores investimentos se concetravam na representação (17,7%) e na aquisição (16,5%) do conhecimento (COELHO, 1990).

2.2 - A IMPORTÂNCIA DO RECONHECIMENTO DE FALA

De qualquer forma, a grande alavanca para erguer novamente o interesse pelo reconhecimento de fala é a real possibilidade de que isso, de fato, aconteça. Apesar dos contrapesos, essa possibilidade está cada vez mais próxima do cotidiano humano. Novas técnicas tem feito com que estas idéias tenham cada vez mais uma aplicação confiável, tornando-as, assim, com grandes possibilidades de comercialização. Além do Voice Pilot da Microsoft, outros produtos lançados comercialmente têm tido algum destaque como

reconhecedores de fala. O Voice Blaster lançado em 1993 pela COVOX (VOICE, 1993), e o VOICETYPE em 1994 pela IBM (VOICETYPE, 1994) são exemplos dessa tendência do mercado.

Essa tendência é também assinalada por PENZIA (Apud MOON, 1995) quando profetiza que no futuro os telefones não terão teclas, serão acionados por comandos de voz. Diversas revistas especializadas em informática apontam essa tecnologia como uma meta a ser alcançada nos próximos anos. Recentemente (setembro de 1995), a revista Byte (AS 20 PRINCIPAIS, 1995), por exemplo, publicou as 20 principais tecnologias para o ano 2000, entre elas, o reconhecimento de fala. Destaques como esse estão se tornando freqüentes em revistas, jornais e programas de TV, o que tudo indica é que logo deverão ser freqüentes, também, no nosso cotidiano.

2.3 - BREVE HISTÓRIA DO RECONHECIMENTO DE FALA

O reconhecimento de fala surgiu recentemente como um atrativo para sistemas computadorizados. A sua história está inteiramente relacionada com a história dos microcomputadores e da evolução do processamento do sinal digital. LUCE (1989) inclusive descreve, muito suscintamente, os passos necessários para a implementação de um reconhecedor. Nessa descrição, LUCE não menciona a técnica utilizada para o tratamento do sinal digital, ou como este deverá ser avaliado para deduzir que palavra foi falada. Sua ênfase se situa, principalmente, na transdução do sinal de voz, ou seja, sua conversão de som para sinal digital. Uma vez digitalizado o som, LUCE propõe uma memória de vocábulos que, comparados com o sinal devidamente processado, disponibilizaria a palavra corretamente escrita para o usuário.

Essa visão do reconhecimento de fala foi muito difundida no começo e, assim, essa tecnologia foi tratada como de fácil dominação. A medida em que se descobria as diferenças de frequência e intensidade da voz de uma pessoa para outra, descobria-se, também, que já não era tão fácil reconhecer a fala humana.

De fato os experimentos com sucesso datam do final da década passada (final de 80) e do início da década atual (90). Os estudos realizados na década de 70 foram isolados e se fixaram mais pelo estudo da fala, da fonética e das possíveis técnicas para produzir o tão desejado reconhecimento. Desta década, a de 70, podemos citar trabalhos de base como o de NEWELL et alii (1973) e DIXON e MARTIN (1979), onde desenvolveram estudos preliminares a respeito do som, e do reconhecimento das palavras sob o ponto de vista de sistemas automáticos.

Ainda na década de 70, houveram diversas tentativas no sentido de aprofundar os estudos no reconhecimento de fala. REDDY (1975), através de uma revista da IEEE, publicou uma série de artigos acadêmicos a respeito do tema. Idéia seguida por HATON (1982), que publicou um livro intitulado "*Automatic Speech Analysis and Recognition*". Entretanto, muitas das propostas eram quase que impraticáveis, nem tanto pelas idéias, mas pela tecnologia disponível na época e pelo seu difícil acesso.

Os anos foram passando e muito problemas tecnológicos foram sendo superados. No final da década de

80, o mundo, finalmente, conheceria os microcomputadores, os disquetes, os winchesters e, também, a conversão analógico-digital, tecnologia proporcionada por placas dedicadas e recurso essencial para o processamento do sinal de som digitalizado. Essa revolução tecnológica acabou afetando muitas outras áreas da computação, e, por tabela, outras ciências de interesse.

Essa revolução não foi diferente para o reconhecimento de fala. A globalização e a popularização dos computadores aumentou significativamente o número de pessoas pesquisando técnicas e tecnologias diversas. Contudo, apesar de toda a tecnologia disponível, um dos maiores trunfos ainda estava por vir. O ressurgimento das redes neurais artificias no início da década de 80 deu um novo impulso à tecnologia do reconhecimento de padrões (TECNOLOGIA I, 1989).

Até então, o modelo mais utilizado para reconhecer a fala era o *Hidden Markov Model*, conhecido como HMM (LANG, WAIBEL e HINTON, 1992, HAYKIN, 1994 e NÁDAS, 1994). Segundo LANG, WAIBEL e HINTON (1992), apesar de ser um modelo de reconhecimento de voz simples, comparado com o conhecimento acumulado de especialistas em fonética acústica, mostrou o quanto é difícil formalizar o conhecimento desses especialistas em forma de um algorítmo de reconhecimento de fala. Assim, por mais simplista que fosse o modelo HMM, foi o mais usado para essa prática, descartando a utilização de sistemas especialistas para essa tarefa.

O modelo de processamento HMM foi extensivamente usado pela IBM sendo substituído por redes neurais no final dos anos 80, quando as redes apresentaram uma maior eficiência em reconhecer as palavras "*bee*", "*dee*", "*ee*" e "*vee*" (chamado de vocabulário confuso BDEV). Essas palavras são difíceis de distinguir porque possuem curta duração, baixa intensidade vocal e, praticamente, quase a mesma frequência. Sem entrar em detalhes de aquisição e processamento do sinal, o sistema da IBM, funcionando com o modelo HMM padrão, possuía uma performance de 80 %. Realizando experimentos com redes neurais, os cientistas da IBM treinaram uma rede para reconhecer as mesmas 4 palavras, e obtiveram uma performance de 90.9 % (LANG, WAIBEL e HINTON, 1992) a mais próxima da performance humana de reconhecimento. A IBM, na época, realizou, também, testes com as mesmas 4 palavras com seres humanos, a performance humana foi de 94 %.

Segundo NÁDAS (1994), nessa mesma época, avaliando a performance das redes neurais, não tardou a aparecer trabalhos sugerindo a combinação das redes neurais com o modelo HMM. Essa mesma realidade também foi observada por LANG, WAIBEL e HINTON (1992).

As redes neurais artificiais se baseiam na neurotransmissão ocorrida no sistema nervoso dos animais para lançar suas bases de fundamentação. Nesta linha, faz-se uma analogia entre células nervosas vivas e o processo eletrônico. Enfatizando o aprendizado dos sistemas como forma de captação de conhecimento (ALEKSANDER, 1990), esta técnica vem sendo extensivamente utilizada para o reconhecimento de padrões como fala e visão (TECNOLOGIA I, 1989).

Havendo tecnologia, tanto de hardware quanto de software, pouco bastava para que as idéias fossem colocadas em prática, era apenas uma questão de tempo. Logo, no final da década de 80, artigos sobre

reconhecedores de fala já começavam a aparecer nos congressos de computação.

2.4 - O RECONHECIMENTO DE FALA E REDES NEURAIS ARTIFICIAIS

As redes neurais artificias, sem dúvida alguma, alavancaram o processo do reconhecimento de fala. Muitas pesquisas têm obtido algum êxito, e, dessa forma, tem feito muitos cientistas migrar para essa tecnologia.

Citando um trabalho que tem obtido um voto de confiança internacional é o *"Neural" Phonetic Typewriter* do pesquisador finlândes TEUVO KOHONEN (1992). Esse trabalho, publicado originalmente em 1988 na IEEE *computer magazine*, demonstra uma pesquisa realizada inteiramente por KOHONEN sobre o reconhecimento de fala. A rede neural possui uma estrutura de forma a reconhecer fonemas, chamando-se, assim, de mapa fonético.

No caso de o reconhecedor de fala identificar palavras, outro trabalho conhecido é o reconhecedor de palavras isoladas de ALEX WAIBEL, KEVIN LANG e GEOFFREY HINTON (1992). Neste trabalho desenvolvido no ano de 1990, WAIBEL, LANG e HINTON usam uma rede neural tipo *back-propagation* para reconhecer o vocabulário confuso BDEV (as palavras inglesas "*bee*", "*dee*", "*ee*" e "*vee*") falado por 100 locutores com uma taxa de acerto de até 97 %.

BART KOSKO, em 1992, também realizou testes com uma rede neural para reconhecimento de fonemas. Os fonemas treinados foram as cinco vogais /a,e,i,o,u/, duas fricativas /f,s/, uma nasal /n/ e outra com som explosivo /t/. Sua técnica utilizou um pré-processamento pela Transformada de Fourier.

Em 1993, MARCEL HUGO e PAULO LUNA (1993) desenvolveram uma aplicação para o reconhecimento de fala utilizando o modelo proposto por KOHONEN (1987). Obtiveram resultados com margem de acerto de 76 % até 98 %, dependendo da palavra pronunciada. O conjunto de treinamento não foi composto por fonemas (como sugere o modelo original de KOHONEN), mas por palavras (a exemplo de WAIBEL, LANG e HINTON) : cadastros, lançamentos, cálculo, emissão, manutenção, outros, fim. HUGO, em 1994, procurou utilizar a Transformada de Fourier como técnica de pré-processamento do sinal de som, entretanto, não recomenda a sua utilização uma vez ter obtido taxas maiores de acerto sem esse tipo de processamento.

Nesse mesmo ano, 1993, RAHIM (1994), pesquisador da *AT&T Bell Laboratories*, publicou um trabalho que descreve seus experimentos com uma rede neural para reconhecimento de fonemas. Suas pesquisas trabalharam com 36 fonemas da língua inglesa extraídos de 462 locutores diferentes. RAHIM (1994) utilizou neste trabalhou mais de 33000 padrões para formar o conjunto de treinamento da rede neural. Seus resultados são tão bons quanto ruins. Bons quando a taxa de acerto para vogais e ditongos alcança 82.9 %, e ruins quando a taxa de acerto das fricativas atinge apenas 38.6 %.

Em todo o caso, o exemplo de reconhecimento de fala mais bem sucedido talvez seja da empresa IBM. A empresa vem pesquisando esse tipo de tecnologia há mais de 20, antes mesmo do ressurgimento das

redes neurais artificiais. Atualmente a empresa conta com o mais bem sucedido reconhecedor de fala disponível no mercado, o VoiceType Dictation. Desenvolvido pela IBM em conjunto com o Centro de Pesquisa Thomas J. Watson, o VoiceType é um sistema do tipo *speaker-dependent* (depende-do-locutor) que captura palavras isoladas (VOICETYPE, 1994). Segundo divulgação da própria empresa, esse sistema possui um vocabulário de até 22000 verbetes com capacidade para converter de 70 a 100 palavras por minuto.

2.5 - OS PROBLEMAS DO RECONHECIMENTO DE FALA

Para os seres humanos, reconhecer a fala é uma tarefa simples e bastante natural. Não pode-se dizer o mesmo dos computadores. Fazer um computador responder a um comando falado é uma tarefa extremamente difícil e complexa. Muitos são os fatores variantes no reconhecimento da voz humana. WHEDDON (Apud HAYKIN, 1994) classificou um dos problemas como sendo o tamanho do vocabulário. O mesmo fato é apontado por WAIBEL e LEE (1990).

Segundo WAIBEL e LEE (1990), existe ainda a questão da palavra isolada e da palavra concatenada. Essa questão se refere ao fato de como exatamente as palavras estão sendo adquiridas pelo sistema. As palavras faladas isoladamente apresentam um aspecto, quando pronunciadas de forma concatenadas apresentam outro. Esse processo, da palavra concatenada, envolve um sistema de procura da palavra dentro de um seqüência inteira de sons.

Outro problema apontado por WHEDDON (Apud HAYKIN, 1994) é que as palavras raramente são pronunciadas da mesma maneira duas vezes. Essa mudança, por menor que seja, certamente afetará o circuito de decisão estabelecido pelo sistema. Ainda sobre fatores de linguística, existe o problema *speaker-dependet* (depende-do-locutor) versus *speaker-independet* (independe-do-locutor). Esse fator é importante e precisa ser considerado. Um sistema tipo *speaker-dependet*, para ter uma boa performance necessita de um novo treinamento toda vez que mudar o orador do sistema. Os sistemas tipo *speaker-independet* possuem a vantagem de serem treinados para diversos tipos de oradores, entretanto, sua performance é menor que os sistemas *speaker-dependet*.

Segundo LIPPMANN (Apud HAYKIN, 1994), esses fatores todos fazem com que as pesquisas de reconhecimento de fala permaneçam em contínua mudança. São esses os fatores que, por ora, impedem que se produza um sistema comercialmente perfeito e 100 % confiável.

Lembramos, porém, que os problemas apontados acima não incluem, em momento algum, processamento ou análise de sinal, seja analógico ou digital. Muitos outros problemas podem ser encontrados dependendo da técnica utilizada no processamento do sinal. O sinal, na maioria das vezes, precisa e deve sofrer algum tipo de tratamento (modificação), de modo a facilitar o seu uso e análise. Para realizar esses tratamentos, é necessário um conhecimento mais apurado sobre o som e a voz humana de uma maneira geral. A fase de aquisição de fala e processamento do sinal, impreterivelmente, antecede a fase de análise da voz.

BACK CHAPTER GO TO INDEX NEXT CHAPTER