Reconhecimento de Voz para Comandos de

Direcionamento por meio de Redes Neurais

João Francisco Valiati

Paulo Martins Engel **(orientador)**

jvaliati@inf.ufrgs.br

Resumo

A constante busca para aperfeiçoar e estreitar o relacionamento entre homens e máquinas, tornando-o mais natural, não é nenhuma novidade. Consequentemente, o reconhecimento de voz possibilitará uma manipulação mais fácil e prática de equipamentos providos com a capacidade de compreender a fala humana. Graças aos avanços tecnológicos na área de hardware esta tarefa torna-se cada vez mais possível de ser realizada. Diante desta realidade, este trabalho propõe a construção de uma interface em computador que reconheça um vocabulário de palavras de forma isolada e independente do locutor, composto por comandos de direcionamento aplicados a um veículo autônomo, utilizando técnicas de processamento de sinais e redes neurais no processo de reconhecimento.

1. INTRODUÇÃO

Os crescentes avanços tecnológicos fazem com que o reconhecimento de fala seja um campo de estudos fascinante e ao mesmo tempo desafiador, uma vez que é bastante grande a gama de aplicações, onde o uso da voz tem o papel de agilizar e facilitar a realização de tarefas cotidianas, buscando extrair da fala as informações relevantes para a realização do reconhecimento.

Um antigo desejo do homem foi sempre poder controlar suas máquinas por meio da fala, uma vez que as pesquisas relacionadas ao reconhecimento de fala iniciaram em meados deste século, ainda sem a utilização do computador, por meio de experimentos mecânicos que simulavam a produção sonora humana[1].

Com o advento da tecnologia que muito evoluiu, as máquinas predominam em quase todos os cenários, desde a conquista espacial

até às residências. Sendo assim, nada melhor do que dotar tais equipamentos com a capacidade de percepção e compreensão da voz humana, que é a forma mais simples, natural e eficaz do ser humano expressar seus pensamentos, e desta forma humanizar mais o relacionamento homem-máquina.

Atualmente, os estudos pertinentes ao reconhecimento automático de voz (RAV) evoluíram muito graças aos avanços tecnológicos como o computador e as placas de som, as quais tornaram possível a conversão do som em dados digitais. As técnicas de processamento de sinais permitem a extração de características que realmente mereçam destaque, pois atuam no sentido de fornecer não somente a informação de interesse ao processamento de determinada amostra de som, como também ocasionar uma redução considerável na quantidade de informações a serem processadas; reduções de 10 a 100 vezes sobre a informação bruta de uma amostra são consideradas normais. Tais informações serão responsáveis pela produção de padrões a serem identificados numa comparação entre determinada referência registrada e a apresentação de uma nova amostra para teste. Então, o papel do sistema será validar ou não determinada amostra, dependendo do tipo e funcionalidade do sistema de reconhecimento em questão[2,3].

Dentre os vários métodos para comparação entre padrões, onde podemos destacar: o Alinhamento Temporal Dinâmico(*Dynamic Time Warping*, DTW), os Modelos Escondidos de Markov(*Hidden Markov Models*, HMM) e as Redes Neurais Artificiais(*Artificial Neural Networks*, ANN); todos já aplicados em experimentos que pesquisam esta nova tecnologia e que podem ser utilizados individualmente ou em conjunto no processo de reconhecimento[4,5].

Os sistemas RAV podem ser divididos em relação ao trato com palavras em: reconhecimento de palavras isoladas, concatenadas ou contínuas, onde deve ser considerado um intervalo de silêncio no tempo entre a pronúncia de cada palavra a fim de diferenciar esta classificação. Também são consideradas pelo RAV, as pronúncias dependentes de determinado locutor, assim como as que não fazem distinção entre locutores, importando somente o que foi dito e não quem o disse[6].

2. MOTIVAÇÃO

O crescimento nas pesquisas relacionadas ao reconhecimento de voz, juntamente com a variedade de aplicações em que essa nova tecnologia possa vir a ser empregada, tanto na manipulação de sistemas pessoais como para controle de equipamentos e utensílios eletrônicos, favorecem e estimulam a pesquisa neste campo interessante e ao mesmo tempo

desafiador, na exploração dos métodos para processamento de sinais e validação o modelo neural proposto, o qual já se mostra bastante eficaz em problemas de classificação de padrões. Neste trabalho visamos a construção de um sistema capaz de reconhecer comandos de direcionamento pronunciados de forma isolada e independente do locutor aplicados a um veículo autônomo.

3. PROCESSAMENTO DE SINAIS

O processamento de sinais é considerada uma das etapas mais importantes da tarefa de reconhecimento, pois é responsável por extrair do sinal de voz os dados relevantes do sinal e menosprezar a informação redundante, com a finalidade de repassar a informação de interesse à fase de comparação dos padrões, neste caso para treinamento e teste da rede neural.

Vários passos compõem a etapa de processamento de sinais: inicia-se com a aquisição do sinal de voz, onde a informação analógica captada no microfone é convertida em sinais digitais. De todos estes sinais digitais capturados muitos dados são considerados como ruído e informações desnecessárias para a caracterização da amostra. O método para delimitar o início e o fim de uma amostra é a detecção dos seus *endpoints*, ou seja, pontos extremos de uma locução. No reconhecimento de palavras discretas, parte-se do princípio de que a locução é precedida e seguida por um período de ruído de fundo ou silêncio, fazendo-se necessária esta detecção a fim de filtrar o sinal bruto adquirido.

Uma vez realizada a detecção dos *endpoints*, o sinal obtido precisa ser dividido em janelas, as quais representam termos curtos do sinal, onde considera-se que o mesmo seja invariante. A literatura cita que intervalos de 10 à 40 ms atendem a este requisito[4].

Conforme o modelo simples para geração da voz, baseado no mecanismo humano de produção da fala, o sinal da voz é produzido por uma convolução de uma fonte de impulsos, g(t), com a articulação, h(t) referente ao trato vocal, como mostra a figura 1.

Figura 1. Modelo simples de geração de voz(o símbolo ‘*’ representa a convolução)

Por este modelo constata-se que se torna necessária a obtenção dos coeficientes cepstrais, ou seja, a separação dos sinais componentes da fala sobre o modelo simples de geração da fala por meio do procedimento e equações descritas na seqüência do texto, uma vez que a informação cepstral extraída contém os dados relativos ao trato vocal e representam as características do sinal amostrado.

- sabendo-se que o sinal de voz é representado pela equação (1)

  $s(t)=g(t)*h(t)$, (1)
  onde " * " representa a convolução,

- deve-se aplicar a Transformada Discreta de Fourier(DFT) sobre s(t), g(t) e h(t), respectivamente, gerando (2)

  $s(w)=g(w)h(w)$, (2)

- tomando-se o logaritmo dos módulos de (1) é obtida (3)

  $\log|s(w)|= \log|g(w)| + \log|h(w)|$, (3)

- sendo os coeficientes cepstrais(cepstro) de s(t), adquiridos por (4)

  $c(n)=F^{-1} \log|g(w)| + F^{-1} \log|h(w)|$, (4)
  onde F^{-1} representa a Transformada Inversa de Fourier(IDFT) e *n* é o índice.

O cálculo dos coeficientes cepstrais é efetuado sobre cada janela, conforme descrito anteriormente, na forma geral de (5), que representa a Transformada Inversa de Fourier.

$$, 0 \leq n \leq N \quad (5)$$

sendo *N* o número de janelas.
De cada janela são extraídos somente os primeiros coeficientes cepstrais, considera-se de 10 a 20 elementos iniciais, os quais estão diretamente relacionados com a função de transferência do trato vocal, uma vez que estes coeficientes não possuem informações relativas à fonte de excitação g(t), a qual é bastante variável. Com isto, os mesmos são usados para representar os sinais relevantes da voz [5].
Um modelo geral do processo de obtenção dos cepstros é ilustrado na figura 2.

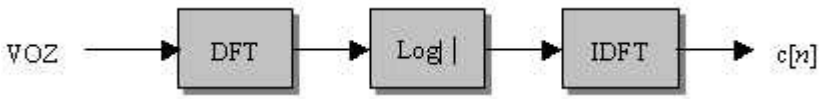

Figura 2. Modelo para obtenção dos cepstros

4. COMPARAÇÃO DE PADRÕES

Com os cepstros extraídos tem-se parâmetros, significativos, a serem apresentados às redes neurais. Primeiramente, antes dos dados serem submetidos a rede neural os mesmos devem sofrer um ajuste temporal no sentido de fixar um tamanho, o qual será referente ao número de entradas da rede. Como a rede a ser utilizada é uma rede *Backpropagation* torna-se necessário o treinamento da rede. Este treinamento é realizado *off-line* e é responsável por estabelecer modelos de referência, que serão atribuídos a determinadas locuções e os parâmetros da rede sofrerão os devidos ajustes.

Após a rede estar devidamente treinada deverão ser apresentadas amostras de teste, onde a identificação de determinada locução é feita em função do padrão de referência que demostre maior similaridade com o padrão de teste.

5. CONCLUSÕES

Através das técnicas de processamento do sinal descritas, visa-se obter informações relevantes a fim de apresentá-las a uma rede neural, que será responsável pela classificação de amostras de teste apresentadas, comparando os padrões já aprendidos com o padrão de teste sugerido. Desta forma procura-se efetivar a identificação de comandos de direcionamento independentes do locutor por meio das técnicas descritas, tornando possível o controle de um veículo autônomo a nível de simulação em computador.

Com isto procura-se reforçar a capacidade de aprendizado das redes neurais artificiais e contribuir na tarefa de reconhecimento de voz, enriquecendo as pesquisas na área de reconhecimento de voz assim como demostrar a efetividade das redes neurais em aplicações desta natureza.

AGRADECIMENTOS

Agradeço ao CNPq pela bolsa de estudos concedida e que torna a realização deste trabalho possível, ao meu orientador professor Paulo Martins Engel, a minha família e a todas as pessoas que tem contribuído para realização deste trabalho.

REFERÊNCIAS

[1] I.H.Witten. Principles of computer speech. New York, Academic Press, 1982.

[2] A.L.Trindade. Estudo de técnicas de processamento de sinais para a geração de gráficos espectrais de sinais digitais. Porto Alegre: CPGCC da UFRGS, 1994.

[3] S. Furui. Digital speech processing, synthesis, and recognition. New York, Marcel Dekker, INC, 1989.

[4] J.G.Proakis. J.R.Deller. J.H.L.Hassen. Discrete-time processing of speech signals. New Jersey, Prentice-Hall,1993.

[5] L.R.Rabiner. R.N.Schafer. Digital processing of speech signal. Prentice-Hall Signal Processing Series, Bell Laboratories, 1978.

[6] P.Foster. T.Schalk. Speech Recognition The Complete Practical Reference Guide, New York, Telecom Library,1993.