IMPLEMENTAÇÃO EM TEMPO REAL DE UM SISTEMA DE RECONHECIMENTO DE DÍGITOS CONECTADOS

Rodrigo V. Andreão e Luis G. P. Meloni

Departamento de Comunicações
Faculdade de Engenharia Elétrica e de Computação – UNICAMP
P.O. Box 6101, CEP: 13083-970 – Campinas – SP – BRASIL
varejao@decom.fee.unicamp.br , meloni@decom.fee.unicamp.br

RESUMO

Este trabalho apresenta um sistema de reconhecimento de dígitos conectados e sua implementação em tempo real. A consistência do sistema é verificada através de testes usando duas bases de dígitos conectados: uma em Português brasileiro e outra em Inglês americano. O sistema é independente de locutor, utiliza modelos ocultos de Markov discretos, modelos de quinze estados por palavras e modelagem de duração de palavras como pós-processamento. Nos testes de reconhecimento, a taxa de acerto de palavras nas simulações ficou ao redor de 99%. O teste do sistema em tempo real, em ambiente de escritório e com uma dezena de locutores, mostrou uma taxa de acerto de palavras de 95,70%.

1. INTRODUÇÃO

As pesquisas em reconhecimento de fala vêm sendo desenvolvidas intensamente pela comunidade científica há mais de 30 anos, e os frutos desse investimento científico começam a ser colhidos.

No início das pesquisas, os sistemas de reconhecimento eram testados através de simulações, onde as condições eram sempre as mais favoráveis, longe do que se encontra em um sistema prático de reconhecimento de fala. Passados anos de pesquisa, somados aos avanços tecnológicos, uma nova realidade se abre ao público em geral com o aparecimento de sistemas de reconhecimento de fala incorporados aos serviços voltados aos usuários.

Os sistemas práticos de reconhecimento de fala contínua são conhecidos por exigir grandes vocabulários, alguns chegando a 250 mil palavras. O desempenho de tais sistemas depende muito da aplicação e do usuário que vai utilizar o serviço. Algumas dessas aplicações requerem o treinamento do sistema por parte do usuário para que sejam alcançados bons desempenhos.

Por outro lado, pode-se conseguir elevado desempenho dos sistemas de reconhecimento de fala, reduzindo-se a complexidade da aplicação. Estudos recentes indicam que os sistemas de reconhecimento atuais, voltados para aplicações restritas independentes de locutor e que utilizam vocabulários reduzidos, podem atingir taxas de acerto de palavras de até 98% [3].

Os requisitos de desempenho de tais sistemas são ainda mais severos quando se pretende operar em tempo real e num ambiente ruidoso.

2. CARACTERÍSTICAS DO SISTEMA DE RECONHECIMENTO DE DÍGITOS CONECTADOS IMPLEMENTADO

Os sistemas automáticos de reconhecimento de fala são baseados em modelos de referência que são obtidos a partir do emprego de bases de sinais. Em geral, essas bases são divididas em duas, uma para o treinamento dos modelos e outra para teste. O processamento envolve a conversão das formas-de-ondas em seqüências de vetores acústicos. Estes vetores são uma representação compacta do conteúdo espectral do sinal cobrindo períodos típicos de 10 ms. Na modelagem realizada neste trabalho, estes vetores são quantizados de forma vetorial. A partir destes vetores são construídos modelos ocultos de Markov discretos, que podem basear-se em modelos de sub-unidades de palavras ou modelos de palavras.

O sistema de reconhecimento se divide em:

- *Pré-processamento* – conta com as etapas de análise espectral, responsável pela parametrização do sinal de fala, e quantização vetorial, que classifica os vetores de parâmetros gerados pela análise espectral.
- *Decodificação* - responsável pela escolha da seqüência de dígitos mais provável. A decodificação é feita pelo algoritmo *One Step* [6], que faz associação de cada dígito a um modelo de referência. Os modelos de referência são os modelos ocultos de Markov discretos.

2.1 Análise Espectral

Levando-se em conta a possibilidade de simular o sistema de reconhecimento como um serviço de telefonia, adotou-se a freqüência de trabalho de 8 kHz. As bases de sinais empregadas nas simulações foram ajustadas para esta freqüência de amostragem. A filtragem é feita por um filtro passa-banda comum na telefonia digital.

Todo o processamento feito no sinal de fala pode ser resumido nos seguintes passos:

- Filtragem de pré-ênfase com fator 0.95;

- Aplicação de janelas de Hamming de 20 ms com superposição de 50%;
- Cálculo do espectro de freqüência via FFT (Transformada Rápida de Fourier) com a aplicação de um banco de filtros triangulares na escala mel;
- Extração de 12 coeficientes cepstrais e energia do espectro de freqüência de cada quadro.

O vetor de parâmetros de dimensão 39 é formado pelos coeficientes cepstrais, energia e suas derivadas primeira e segunda.

2.2 Quantização Vetorial

Os *codebooks* são gerados a partir da base de sinais de treinamento utilizando o algoritmo LGB [5].

São empregados *codebooks* independentes para cada parâmetro. O tamanho dos *codebooks* utilizados é de 256 vetores para os parâmetros cepstrais e de 32 vetores para as energias.

O treinamento dos vetores dos *codebooks* emprega a medida de distorção Euclidiana e o critério de convergência é a distorção menor que 1%.

2.3 Modelo Oculto de Markov Discreto – HMM (do inglês, *Hidden Markov Models*)

Para o caso de vocabulários pequenos, como o de dígitos conectados, é mais interessante trabalhar com HMM de palavra ao invés de sub-unidades fonéticas. Desta forma, necessita-se de um número inferior de modelos. Assim, optou-se pelo HMM de palavras com estrutura *left-right*, exemplificado na Figura 1.

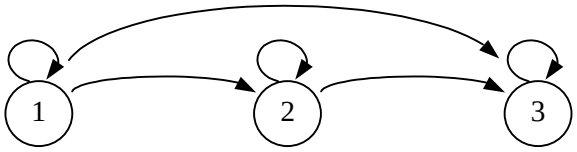

Figura 1 Estrutura de HMM de palavra com 3 três transições por estado.

O sistema emprega quinze estados para todos os modelos, exceto o modelo de silêncio que usa apenas dois estados.

2.4 Decodificação: Algoritmo *One Step*

O algoritmo *One Step* [6][10] faz uma busca síncrona no tempo da seqüência de dígitos mais provável. O número de níveis máximo é um parâmetro de entrada do algoritmo. Cada nível representa um dígito. Como resultado da busca, o algoritmo retorna a melhor seqüência de dígitos.

A seqüência de dígitos pode possuir tamanhos variados. Assim, o algoritmo retornará as melhores seqüências de dígitos até o máximo definido; e, dentre estas, a melhor. Emprega-se um número máximo de níveis igual a 10.

2.4.1 Incorporação de Modelo de Duração

O modelo de duração é necessário para dar maior consistência ao sistema de reconhecimento. O sistema emprega um modelo de duração de palavras [9].

O modelo de duração de palavras é parametrizado por uma função de densidade de probabilidade normal caracterizada pela equação:

$$p_n(d) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(d-\mu_n)^2}{2\sigma_n^2} \right) \quad (1)$$

onde μ_n é a duração média da palavra *n*, σ_n é a variância e *d* é a duração. Todos os dados de duração são expressos em números de quadros.

A probabilidade de duração da palavra é acrescentada, pelo algoritmo de busca, num pós-processamento e, assim, permite diferenciar melhor as palavras pela sua duração.

3. BASES DE SINAIS

Foram usadas duas bases de sinais para avaliação do desempenho do sistema. Uma do Laboratório de Processamento Digital de Fala que foi denominada neste trabalho de LPDF Dígitos e outra em Inglês americano chamada de TI *Digits*.

3.1 LPDF Dígitos

A LPDF Dígitos [4] é uma pequena base de dígitos conectados em Português gravada na freqüência de amostragem de 11.025 kHz. Ela foi dizimada para a freqüência de 8 kHz. O ambiente de gravação é o de escritório e feito através de placas de som de computadores pessoais.

O vocabulário desta base compreende todos os dígitos de 0 a 9 e mais a palavra *meia*.

A base é dividida num conjunto de treinamento e de teste como mostrado na Tab. I. O conjunto de treinamento possui frases com seqüências de 8 dígitos e estas são diferentes das frases usadas no conjunto de teste. Os locutores também foram separados para possibilitar a condição de independência de locutor.

Tabela I Base de sinais LPDF Dígitos

Conjunto	Locutores	
	Femininos	Masculinos
Treinamento	13	18
Teste	4	5

Cada locutor pronunciou 11 frases escolhidas aleatoriamente de um conjunto de 55 frases diferentes. São permitidas repetições entre as frases para um mesmo locutor. Assim, o conjunto de treinamento totaliza 341 frases e o conjunto de teste 99 frases.

3.2 TI *Digits*

A TI *Digits* [7] é uma base de dígitos em Inglês americano. Foi gravada na freqüência de 20 kHz. Esta base é também dizimada para a freqüência de 8 kHz. Esta base toma cuidados especiais no que diz respeito ao ruído ambiente e ao equipamento de gravação.

O vocabulário desta base compreende os dígitos de 0 a 9 e mais a palavra *oh* (zero).

A base possui um conjunto de treinamento formado por 112 locutores, sendo 13 femininos e 18 masculinos. Cada locutor pronunciou 11 frases escolhidas aleatoriamente de um conjunto de 55 frases diferentes. Assim, o conjunto de treinamento totaliza 341 frases. Não são permitidas repetições entre as frases para um mesmo locutor.

A TI *Digits* é completa no sentido que possui seqüências de dígitos de vários tamanhos, desde 1 até 7 dígitos conectados. A distribuição do número dos conjuntos de treinamento e teste é mostrada na Tab. II.

Tabela II Base de sinais TI *Digits*

Conjunto	Locutores	
	Femininos	Masculinos
Treinamento	57	55
Teste	57	56

Cada locutor pronunciou 77 seqüências de dígitos geradas aleatoriamente. Essas seqüências foram divididas nas seguintes categorias:

- 22 seqüências de dígitos isolados, sendo duas repetições de cada um dos 11 dígitos do vocabulário;
- 11 seqüências de dois dígitos;
- 11 seqüências de três dígitos;
- 11 seqüências de quatro dígitos;
- 11 seqüências de cinco dígitos;
- 11 seqüências de sete dígitos.

O conjunto de treinamento totaliza 8622 frases e o conjunto de teste 8700 frases.

4. TREINAMENTO DOS HMM's

Na etapa de treinamento são estimados os modelos HMM referente a cada palavra. A estimação dos HMM's requer um conjunto representativo de elocuções que permita uma boa estimativa de cada modelo. A tarefa de segmentação da seqüência de dígitos deve ser bem executada, pois influenciará no desempenho do sistema.

A segmentação automática apresentada em [9] requer uma boa iniciação dos HMM's. Essa exigência é ainda maior quando se trabalha com modelos discretos [10]. Por isso, os modelos são iniciados por um pequeno conjunto de treinamento de dígitos isolados. O conjunto de dígitos isolados é obtido segmentando-se manualmente um conjunto pequeno de elocuções (seis) para cada dígito do conjunto de treinamento da base LPDF Dígitos, igualmente divido entre locutores masculinos e femininos. A iniciação consiste no treinamento dos modelos via algoritmo Baum-Welch. Com os modelos iniciados, realiza-se o treinamento de palavras conectadas descrito no diagrama de blocos da Figura 2.

Os *Arquivos de Treinamento* compostos por dígitos conectados são segmentados em dígitos isolados (*Elocuções*) pelo algoritmo de *Segmentação Automática*. A *Segmentação Automática* é feita pelo algoritmo *Level Building* [9][10], que determinará a fronteira entre cada dígito. A particularidade do treinamento é o conhecimento a priori da seqüência correta de dígitos.

As *Elocuções* obtidas de cada arquivo são armazenadas em recipientes associados ao HMM correspondente. Depois que todos os *Arquivos de Treinamento* são segmentados, é executado o *Treinamento* propriamente dito.

O *Treinamento* de cada HMM é feito através do algoritmo Baum-Welch [8][9][10]. Nesta fase, são obtidas as informações para o modelo de duração de palavra.

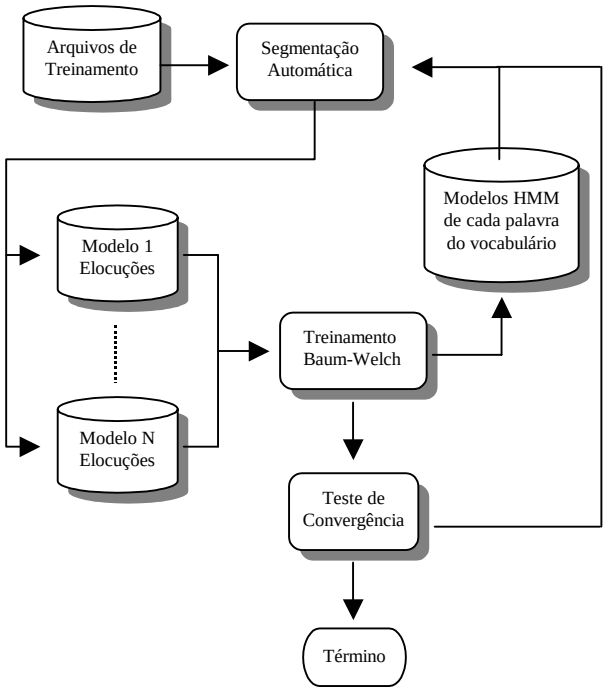

Figura 2 Diagrama de blocos do sistema de treinamento

Terminado o *Treinamento* de todos os modelos, o teste de convergência é executado. A convergência ocorre quando a distância entre os modelos antes e depois de uma época de treinamento for menor que 0.1% [10]. A medida de distância é dada pela normalização da verossimilhança média de todas as elocuções de treinamento de um determinado modelo.

5. IMPLEMENTAÇÃO EM TEMPO REAL

A implementação em tempo real de um sistema de reconhecimento de fala levanta algumas questões importantes:

- Custo computacional;
- *Hardware* adequado;
- Adaptar o sistema de reconhecimento para o processamento em tempo real.

A questão do custo computacional é avaliada como a capacidade que o sistema tem para realizar os processamentos necessários dentro de um intervalo de tempo pré-determinado. Esse intervalo de tempo está associado ao período de atualização do quadro, considerando-se que o sinal de fala é divido em quadros de 20 ms com superposição de 50 %, um período adequado é de 10 ms. Por outro lado, a escolha do *hardware* pode ajudar na questão de custo computacional. Inicialmente, o sistema foi testado num microcomputador PENTIUM III 500 MHz, conseguindo-se a operação em tempo real.

O sistema para operar em tempo real deve estar sincronizado no tempo. O diagrama de blocos do sistema adaptado ao processamento em tempo real é apresentado na Figura 3.

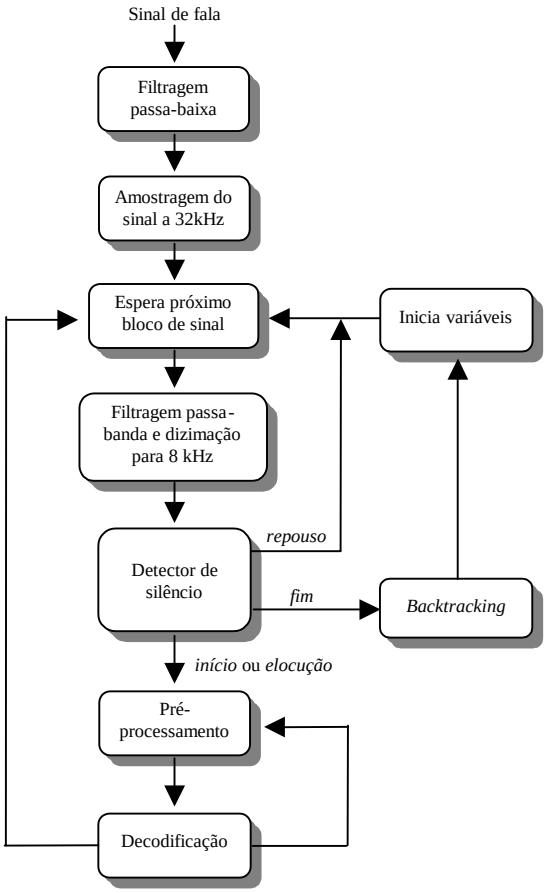

Figura 3 Diagrama de blocos do sistema de reconhecimento de fala.

As etapas iniciais de filtragem passa-baixa e amostragem a 32 kHz são executados pela placa de som do microcomputador. Em seguida, as amostras do sinal são armazenadas num bloco de 200 ms. O bloco de 200 ms é processado por um filtro passa-banda (300-3400 Hz) e dizimado a 8 kHz. Estes dois últimos processamentos visam simular as características em freqüência do canal telefônico.

O bloco de sinal dizimado a 8 kHz é enviado ao detector de silêncio. O detector de silêncio verificará a presença de sinal de fala. Caso não seja detectada uma elocução, o sistema entra no estado de *repouso*. Por outro lado, se for detectado o início de uma elocução, o sistema entra no estado *elocução* e inicia os procedimentos de pré-processamento e decodificação. Finalmente, se for detectado o final de uma elocução, o sistema entra no estado de *fim*, executa o procedimento de *backtracking* para a recuperação da seqüência de dígitos reconhecida e inicia as variáveis do sistema. Cada estado do detector leva o sistema para a espera de um novo bloco de sinal.

5.1 Detecção de Silêncio

A detecção de silêncio é realizada através de um detector de *endpoint*. Esse detector desempenha algumas funções fundamentais no processamento em tempo real. Entre elas podemos destacar:

- Remoção de períodos longos de silêncio;
- Redução do processamento;
- Descartar falsas elocuções.

O processo de detecção é feito através de duas medidas: taxa de cruzamentos de zero e energia. Essas informações são extraídas a cada 10 ms de um bloco de sinal de 200 ms. Portanto, para cada bloco de sinal recebido, geram-se dois vetores de dimensão 20 para armazenar a taxa de cruzamento de zeros e a energia. Em seguida, são gerados os limiares de detecção. Comparando-se os limiares com a taxa de cruzamento de zeros e a energia, determinam-se pontos limitantes, que são as marcações do quadro inicial e final dentro de cada bloco de sinal de 200 ms.

Há a necessidade de se adaptar os limiares às condições de operação. As condições de operação dependem do equipamento utilizado e do ambiente de trabalho. Por isso, uma vez ativado o sistema de reconhecimento, são extraídos do primeiro bloco de 100 ms de sinal, a taxa de cruzamentos de zero e a energia do ruído. Esses valores são refletidos para um período de 10 ms de sinal, dividindo-se por 10 a taxa de cruzamentos de zero e a energia. O conhecimento dessas variáveis é fundamental para a geração dos limiares do detector.

5.2 Limitações da Detecção de Silêncio

O detector de *endpoint* pode provocar erros de reconhecimento. Naturalmente, sabe-se que um período de silêncio seguinte a uma elocução indica o término do processamento e o início do procedimento de *backtracking*. Não se pode afirmar com certeza que qualquer período de silêncio detectado durante o processamento de uma palavra seja o fim da elocução. A ocorrência de consoantes oclusivas numa palavra gera um pequeno período de oclusão. Isso conduz a uma falsa detecção de silêncio e, por esse motivo, o algoritmo baseia-se em uma máquina de estados.

A Figura 4 mostra uma situação típica onde uma falsa detecção causará uma possível falha no reconhecimento. As barras

verticais no sinal indicam os intervalos de 200 ms de cada bloco do sinal. As barras indicadas pelas setas foram inseridas pelo detector de *endpoint*. O pequeno período de silêncio é detectado pelo algoritmo de *endpoint*, mas não é levado em considerção para evitar a separação das sílabas da palavra "oito".

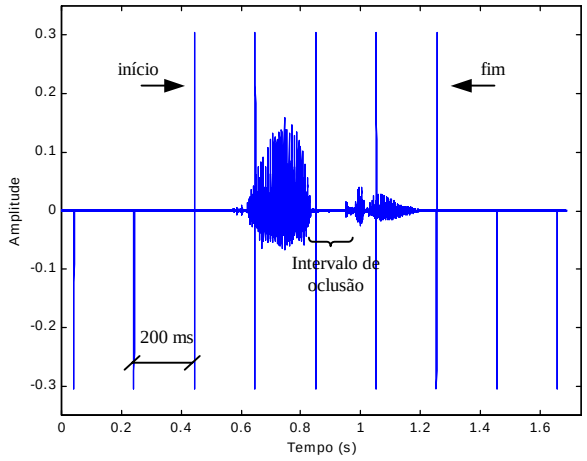

Figura 4 Resultado da detecção da palavra "quatro".

A máquina de estados da Figura 5 adicionada ao detector de *endpoint* visa lidar com tais situações. Os estados são alcançados depois de respeitadas algumas condições, as quais foram ajustadas experimentalmente.

No estado 0, o sistema está em repouso. Nenhum processamento é realizado no sinal, pois existe somente silêncio. Enquanto isso, o sistema está esperando uma nova elocução.

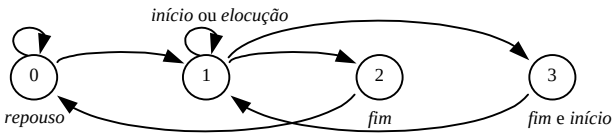

Figura 5 Máquina de 4 estados do detector de *endpoint*.

O estado 1 é atingido quando o início da elocução for detectado. Entretanto, falsas elocuções podem ser descartadas, caso suas durações sejam inferiores a cinco quadros. Por outro lado, se o início da elocução for confirmada, o sistema inicia o processamento do sinal. Deste estado, pode-se transitar tanto para o estado 2, quando ocorrer um período de silêncio no final do bloco superior a 100 ms, quanto para o estado 3, quando a elocução detectada possuir curta duração. Se as condições anteriores não forem atendidas, o sistema permanecerá no estado 1.

O estado 2 é atingido quando for detectado o fim da elocução. Assim, inicia-se o procedimento de *backtracking* e retorna-se ao estado 0.

O estado 3 representa um caso especial de detecção. Ele é atingido quando detectado o fim da elocução e o início de uma nova em um mesmo bloco de 200ms. Deste estado, transita-se diretamente para o estado 1.

6. RESULTADOS

A avaliação de desempenho do sistema foi feita a partir de dois testes: um com as bases de sinais e outro em tempo real. O primeiro teste visa comprovar o desempenho do sistema a partir dos conjuntos de treinamento e teste das bases de sinais utilizadas no trabalho. O teste em tempo real simula uma situação típica de utilização do sistema de reconhecimento de dígitos conectados.

6.1 Teste com as Bases de Sinais

Nestes experimentos são avaliadas as taxas de acerto de palavras e frases, tanto para o conjunto de treinamento quanto para o conjunto de teste. Os resultados obtidos pela base de treinamento são muito úteis, pois permitem avaliar a convergência do algoritmo de treinamento.

Utilizou-se o algoritmo *One Step*. Além disso, acrescenta-se o modelo de duração de palavra como pós-processamento. Não foi feita nenhuma ponderação do modelo de duração de palavra.

Os resultados obtidos com a base TI *Digits* consideraram o número de dígitos na frase conhecido TC e desconhecido TD, objetivando explorar melhor as características da base.

As Tabelas III e IV abaixo apresentam os resultados obtidos com as bases LPDF Dígitos e TI *Digits* respectivamente.

Tabela III Resultados com a base LPDF Dígitos

Conjunto	Acerto de Palavras (%)	Acerto de Frases (%)
Treinamento	99,89	99,12
Teste	99,37	94,95

Tabela IV Resultados com a base TI *Digits*

Conjunto	Acerto de Palavras (%)		Acerto de Frases (%)	
	TC	TD	TC	TD
Treinamento	99,58	99,18	98,74	97,61
Teste	99,31	98,64	97,93	96,15

Comparando-se as taxas de acerto de palavras entre as duas bases, pode-se verificar uma equivalência dos resultados. Além disso, a taxa de acerto de palavras ficou acima de 99% para quase todos os casos.

A utilização de modelo de duração de palavras melhora o desempenho do sistema, e, em último caso, não influencia nos resultados, caso constatado com a base LPDF Dígitos. Além disso, o custo computacional sofre um acréscimo desprezível. Por estas razões, considera-se sua utilização uma boa alternativa

na melhoria de desempenho do sistema de reconhecimento de fala.

As condições TC e TD são situações em que um sistema de reconhecimento de fala pode ser submetido. Se for dada liberdade ao locutor em pronunciar um número indefinido de dígitos, ter-se-á uma queda no desempenho do sistema, como demonstrado pela Tabela IV. Entretanto, sempre que for possível definir o número de dígitos a ser pronunciado pelo locutor, ocorre melhora no desempenho e no grau de confiança do sistema.

6.2 Teste em Tempo Real

O teste em tempo real foi realizado com uma dezena de locutores adultos do sexo masculino. Nenhum desses locutores participou da geração da base de treinamento. Foi solicitada a pronúncia de seqüências de dígitos com naturalidade. Todas as seqüências contêm oito dígitos e representam números de telefones escolhidos pelo próprio locutor. Cada locutor pronunciou quatro números de telefones distintos.

O ambiente de teste é o de escritório. Não foram tomados cuidados com ruído de fundo. O microfone é de eletreto com suporte na cabeça.

Foram utilizados os modelos ocultos de Markov obtidos com a base LPDF Dígitos. A busca feita pelo algoritmo de decodificação simulou frases com número de dígitos variando de um a dez.

A partir dos testes realizados, o sistema apresentou uma taxa de acerto de palavras de 95,70%. Vale ressaltar que os locutores da base de treinamento são, em sua maioria, da região de São Paulo, e os locutores utilizados no teste são de diferentes regiões do país.

Entre os erros encontrados, o mais significativo foi a troca do dígito três pelo dígito seis, representando 82% dos erros. Isso se repetiu até quando foi pedido para o locutor pronunciar isoladamente esse dígito. Necessita-se, portanto, investigar mais o processo de geração do modelo para o dígito três.

7. CONCLUSÃO

Os resultados obtidos neste trabalho foram os melhores obtidos até o presente em nosso laboratório com a base LPDF Dígitos [4]. Um dos motivos para isso é a utilização de um procedimento de treinamento mais adequado, além do aproveitamento de novas informações obtidas através do modelo de duração de palavras. Além disso, os testes com a base comercial TI *Digits* só vieram a confirmar o bom desempenho do sistema.

Espera-se resultados ainda melhores com a utilização de HMM´s contínuos [9] e a adoção de uma metodologia de treinamento discriminativo [2][3], que poderão contribuir nos resultados obtidos neste trabalho.

A implementação em tempo real buscou um compromisso de baixo custo computacional e baixa taxa de erro. Entretanto, o sistema de reconhecimento de fala sofre uma queda significativa de desempenho quando testado num ambiente diferente do de simulação. Essa queda de desempenho é levantada na literatura, mas ainda não foram apresentadas soluções que definitivamente reduzem a grande diferença entre os resultados de simulações e de operção em tempo real.

Necessita-se aumentar a robustez do sistema às interferências externas ao sinal de fala. Além disso, o desempenho do sistema é dependente da qualidade do equipamento de aquisição do sinal de fala; do microfone e da placa de áudio. Por fim, necessita-se prever situações comuns à utilização em tempo real como hesitações do locutor, sopros entre outras.

Enfim, a implementação de um sistema de reconhecimento de fala numa situação mais realista apresenta grandes desafios, exigindo novas pesquisas visando a melhoria de tais sistemas.

8. REFERÊNCIAS

[1] E.R. Buhrke et al., "*Application of Vector Quantized Hidden Markov Modeling to Telephone Network based Connected Digit Recognition*", ICASSP-94, pp. 105-108, 1994.

[2] R. Cardin, Y. Normandin e E. Millien, "*Inter-Word Coarticulation Modeling and MMIE Training for Improved Connected Digit Recognition*", ICASSP-93 Minneapolis, MN, pp. 243-246, Abril 1993.

[3] R. Comerford, J. Makhoul e R. Shwartz, "*The Voice of the Computer is Heard in the Land (and It Listens Too!)*", IEEE Spectrum, Dezembro 1997.

[4] F.L. Figueiredo, "*Segmentação Automática e Treinamento Discriminativo Aplicados a um Sistema de Reconhecimento de Dígitos Conectados*", Dissertação de Mestrado, UNICAMP, Campinas, 1999.

[5] A. Gersho e R.M. Gray, "*Vector Quantization and Signal Compression*", Kluwer Academic Publishers, 1992.

[6] C. Lee e L.R. Rabiner, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, pp. 1649-1658, November 1989.

[7] R.G. Leonard, "*A Database for Speaker-independent Digit Recognition*", Proceedings of the ICASSP - 84 , pp. 42.11.1-4, Março 1984.

[8] L.G.P. Meloni, "*Learning Discrete Hidden Markov Models*", Computer Applications in Engineering Educat., vol. 8, no. 3, pp. 141-149, 2000.

[9] L.R. Rabiner, J.G. Wilpon e F.K. Soong, "*High Performance Connected Digit Recognition Using Hidden Markov Models"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, pp. 1214-1225, August 1989

[10] L.R. Rabiner e B.H. Juang, "*Fundamentals of Speech Recognition*", Prentice Hall, 1993.