Simulação de um Vocoder Digital

Marco Antonio T. Pereira e Fernando Acatauassu G. Ferreira

Este trabalho apresenta os métodos empregados para a simulação de um vocoder LPC-10 utilizando algoritmos de domínio público, sendo que introduzimos algumas modificações na decisão sonoro/surdo que resultaram em sensível melhoria na qualidade de voz sintetizada. Discutimos o "software" e o "hardware" utilizados para a simulação, e finalizamos apresentando os resultados obtidos baseados em um teste de inteligibilidade para a voz sintética, também por nós desenvolvido.

1. Introdução

Este trabalho discute a simulação de um vocoder ("voice coder"), trabalhando à taxa de 2000 bit/s utilizando um sistema com barramento STD [1]. O vocoder simulado utiliza o método de Codificação por Predição Linear (LPC – "Linear Predictive Coding") [2] e a extração do período fundamental ("pitch") e decisão sonoro/surdo usando o algoritmo de Gold e Rabiner [3] e [4] com uma sofisticação, por nós introduzida, no que se refere à decisão sonoro/surdo.

Apresentamos a seguir o modelo utilizado (que inclui algoritmos modificados, mais eficientes), a forma como o mesmo foi simulado, e os resultados obtidos na simulação através da aplicação de um teste de percepção que criamos como uma ferramenta de auxílio para chegarmos a conclusões mais reais.

Concluímos o trabalho com observações sobre perspectivas de melhoria na qualidade de voz utilizando o modelo explorado.

2. Métodos Utilizados

O modelo que examinamos, o LPC-10 (10 pólos), é um sistema que pode ser classificado como um Codificador Paramétrico[1], ou seja, representa a

1. As designações Codificador do Tipo Análise-Síntese e Codificador de Fonte são também utilizadas como sinônimos de Codificador Paramétrico. Em particular, a expressão Codificador de Fonte é muitas vezes empregada com um sentido bem mais amplo.

Os autores são diretores da Kernel Informática Ltda, SIA Trecho 2, lote 1205, 71200, Brasília, DF.

fala por parâmetros correspondentes a um modelo que define o mecanismo de sua produção. Estes parâmetros são transmitidos e serão utilizados na recepção para reconstrução da mesma. Este modelo supõe independência entre a fonte de excitação e o sistema do filtro que reproduz o trato vocal humano. Assim, podemos analisar e sintetizar a função de transferência do trato vocal sem considerar o tipo de excitação que será utilizado. Este modelo é geralmente utilizado até uma taxa de transmissão de no máximo 4800 bit/s. O LPC está descrito em detalhes em [2] e o seu diagrama em blocos simplificado, representativo do mecanismo de produção da fala, está mostrado na **Fig. 1.**

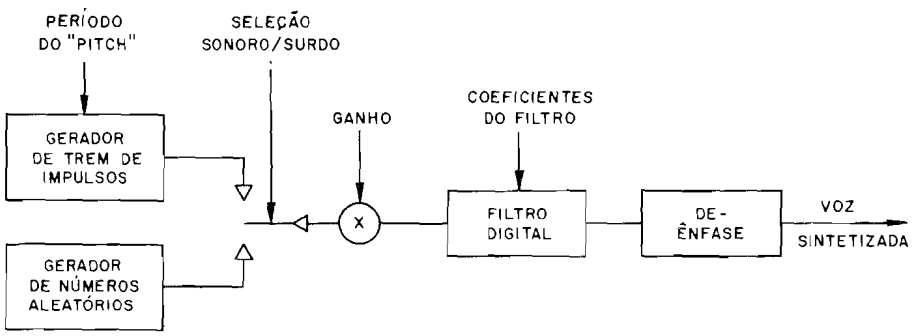

Figura 1. Modelo simplificado para produção da fala.

Conforme esta figura, existem dois tipos possíveis de excitação, ditas sonora e surda. O período fundamental estabelece o período da excitação do gerador de trem de impulsos para excitações tipo sonora, e o gerador de números aleatórios simula a excitação tipo ruidosa dos sons fricativos e oclusivos. O tipo de excitação é selecionado pela chave. Existe um fator de ganho e o filtro digital variante no tempo responsável pelo modelameto da excitação que reproduzirá a fala. A de-ênfase compensa a pré-ênfase, introduzida na análise da voz a fim de compensar os efeitos do espectro do pulso glotal e radiação labial, que resultam em uma queda de cerca de 6 dB/oitava no espectro do sinal de voz.

O conjunto de processos que resultam na extração dos parâmetros correspondentes aos coeficientes de predição e ao ganho G está ilustrado na **Fig.** 2. Após a pré-ênfase, ocorre a extração de parâmetros via autocorrelação, que exige uma janela anterior (Hamming) para reduzir o efeito entre as amostras de janelas consecutivas no processo de análise, e também para limitação do intervalo de análise. Estas janelas, no nosso caso, tem a duração de 25ms e não há superposição entre elas. Para chegar aos parâmetros temos de resolver um sistema de equações, o que é feito utilizando o método de Durbin [2] e [5]. O parâmetro do ganho é extraído a partir dos coeficientes de predição [2]. No caso deste valor estar abaixo de um determinado valor limite, temos um quadro de silêncio. Observe-se que a saída do último bloco

na **Fig. 2** é constituída pelos coeficientes de correlação parcial k_i (i = 1,..., P) e pelo ganho G. A obtenção desse conjunto de parâmetros é equivalente à obtenção dos coeficientes de predição e ganho [2].

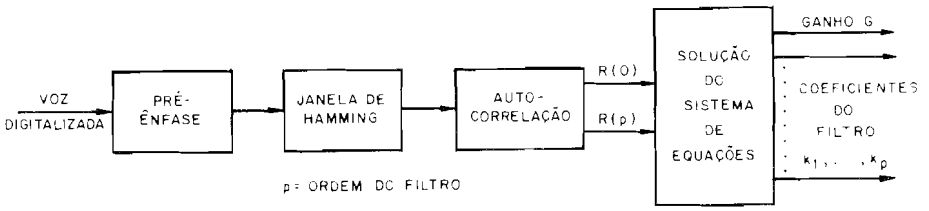

Figura 2. Extração dos coeficientes do filtro e do ganho.

O processo para extração do valor do período fundamental referente ao segmento analisado é iniciado com a aplicação do método de processamento paralelo proposto por Gold e Rabiner [3]. Ele se baseia em uma análise comparativa da amplitude de picos e vales do sinal de voz. São geradas seis medidas diferentes, das quais resultam o valor de duas variáveis que serão utilizadas na determinação do valor final do período fundamental: o período fundamental selecionado pelo algoritmo para um dado quadro analisado e a nota atribuída a este período fundamental no processo de escolha. A nota é um valor ponderado gerado pelo próprio algoritmo de Gold e Rabiner, significando em última análise a probabilidade daquele valor de período fundamental escolhido estar correto.

A princípio a decisão sonoro/surdo estava embutida na própria extração do período fundamental. Um valor de período fundamental zero significa um quadro surdo. Porém, como desta forma a qualidade da voz reproduzida era bastante pobre, a decisão sonoro/surdo, a partir da idéia descrita em [6], foi sofisticada e passou a utilizar algumas outras variáveis. A decisão utiliza agora o valor do período fundamental e nota estabelecidos conforme descrito acima, a taxa de cruzamento de zero no quadro, o valor do ganho (G) já extraído, além do erro médio quadrático. A taxa de cruzamento de zero é o número de vezes que o sinal analógico de voz no quadro sob análise cruzou um valor de referência.

O cálculo do erro médio quadrático utiliza valores derivados da análise LPC. O erro médio quadrático $\tilde{E}_n$ é definido como

$$\tilde{E}_n = \sum_{m=0}^{N-1} e^2_n(m) \qquad (1)$$

onde N é o comprimento da janela e

$$e_n^{(m)} = s_n^{(m)} - \hat{s}_n^{(m)} \qquad (2)$$

com $s_n^{(m)}$ designando a m-ésima amostra na saída da janela de ordem n e $\hat{s}_n^{(m)}$ denotando a predição desta amostra. Representando por E_n o valor mínimo, do erro médio quadrático $\tilde{E}_n$, tem-se então [2]

$$E_n = R_n(0) - \sum_{k=1}^{P} a_k R_n(k) \qquad (3)$$

onde os valores $\{a_k\}$, $k = 1, \ldots, P$ são os coeficientes de predição; P é a ordem do preditor (no caso, 10); e $R_n(k)$ são valores da autocorrelação para a janela considerada, ou seja

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n^{(m)} \; s_n^{(m+k)} \qquad (4)$$

Ao invés de trabalhar com E_n é conveniente definir um erro normalizado $\bar{E}_n$ dado por

$$\bar{E}_n = \frac{E_n}{R_n(0)} \qquad (5)$$

Esta normalização assegura que $0 < \bar{E}_n < 1$ e permite trabalhar em ponto fixo. Na implementação prática, em matemática de ponto fixo, poderíamos incorrer em problemas de precisão numérica não fosse esta precaução.

De posse destas variáveis, o algoritmo de decisão sonoro/surdo estabelece o período fundamental eleito, o qual ou é o valor de período fundamental originalmente extraído pelo algoritmo de Gold e Rabiner, evidenciando um quadro sonoro, ou é zero, evidenciando um quadro surdo. O processo de decisão utiliza os valores das variáveis citadas, que devem se encaixar em faixas de valores predeterminados para as várias tomadas de decisão. Os valores limites destas faixas foram determinados através de experimentação. O fluxograma do algoritmo para tomada de decisão sonoro/surdo está na **Fig. 3**. Este algoritmo, por nós desenvolvido, se aplicou muito bem às situações consideradas. A comparação entre a opção tomada pelo algoritmo com aquela que resultaria de uma inspeção visual do quadro analisado, indica uma precisão de decisão sonoro/surdo de 86%. Para fins de levantamento desta estatística foram efetuadas 50 comparações.

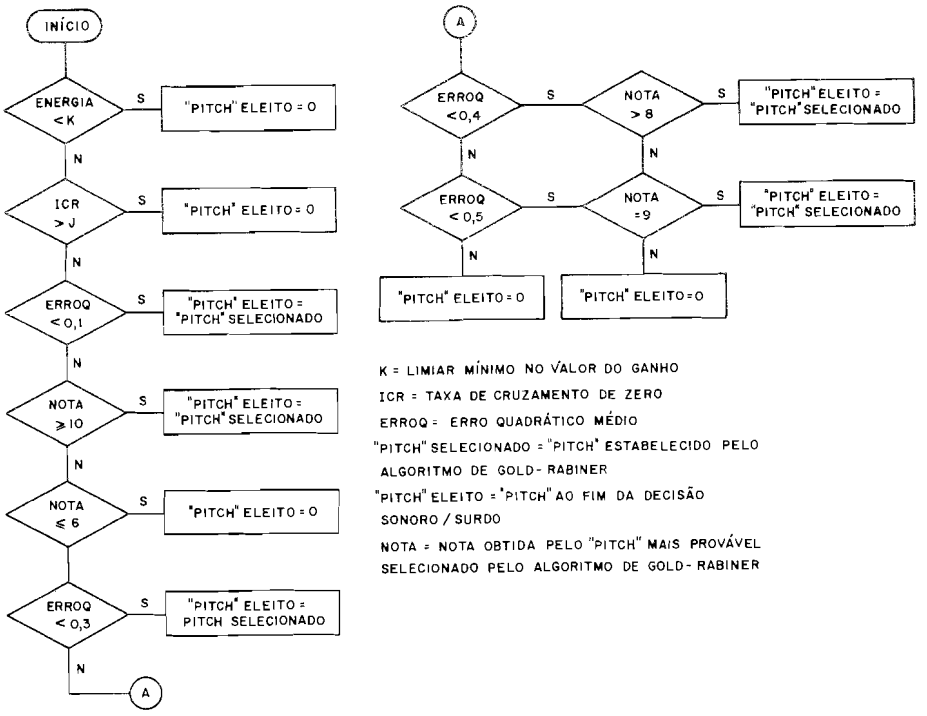

Figura 3. Fluxo para tomada de decisão sonoro/surdo.

Mas este ainda não é o valor utilizado como período fundamental final para o processo de síntese. Existe ainda um algoritmo de suavização para a retirada de erros grosseiros. Ele compara o valor encontrado para o período fundamental com o valor passado e dois adiantados, evitando estes erros. Utilizando os valores do período fundamental destes três quadros e do quadro atual, se em meio a quadros surdos o algoritmo encontra um quadro sonoro, ele automaticamente o transforma em surdo, e vice-versa. Da mesma forma, se no início ou finalização de quadros é detetado um quadro sonoro, ele também é automaticamente transformado em surdo. Desse modo, os erros grosseiros evitados pelo algoritmo de suavização são:

(i) deteção isolada de quadros sonoros ou surdos;
(ii) estimativa de valores de período fundamental diferentes de zero no início ou fim de regiões sonoras.

O diagrama em blocos de todo o processo, a partir das amostras de voz até a obtenção do período fundamental final, está na **Fig. 4**.

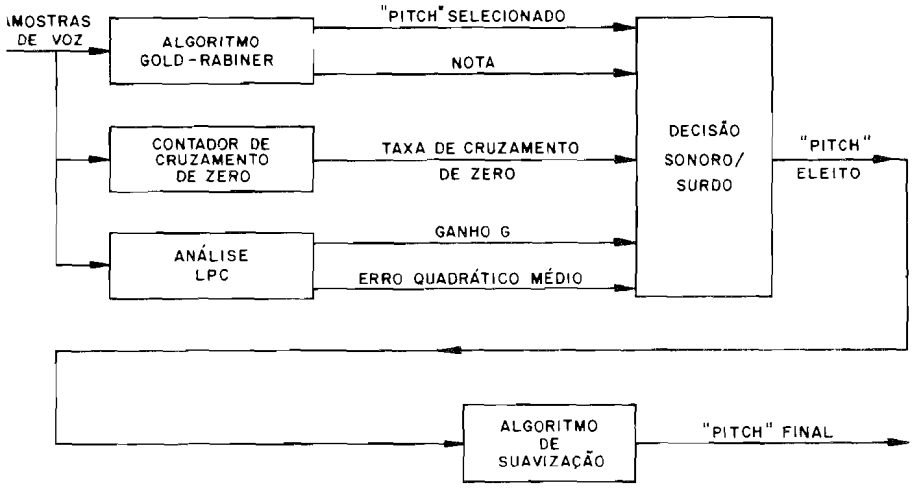

Figura 4. Diagrama em blocos para o processo de extração do "pitch".

Após a extração de todos os parâmetros, eles são codificados da seguinte forma, segundo exigência do sintetizador utilizado [7]: energia, 4 bits; período fundamental, 6 bits, k_1 e k_2, 5 bits cada; k_3, k_4, k_5, k_6 e k_7, 4 bits cada; k_8, k_9 e k_{10}, 3 bits cada; além de 1 bit para indicar repetição de quadro que não foi utilizado. Como a duração da janela utilizada é de 25 ms temos uma taxa de 40 Hz x 50 bits = 2000 bit/s.

3. A Simulação

Para efetuarmos a simulação deste conjunto de algoritmos que resultam em um vocoder, utilizamos um microcomputador baseado no microprocessador 8088 da Intel em um barramento STD. A simulação tem como objetivo fundamental assegurar o funcionamento de cada sub-sistema dentro de uma dada realidade que será a esperada no equipamento final, bem como a integração dos vários sub-sistemas. Entendemos aqui por sub-sistema cada um dos vários módulos de "software". Na simulação, para a observação de desempenho do modelo, podemos criar condições, alterar valores de variáveis e provocar situações de funcionamento nos extremos desejados. Por exemplo, os valores limites para as variáveis utilizadas na decisão sonoro/surdo foram determinados de maneira iterativa nesta fase de simulação, com a ajuda do "hardware" e "software" descritos a seguir.

Para que o microcomputador pudesse obter amostras de voz para trabalhar, e para que depois fosse possível ouvir os resultados, foi desenvolvida uma placa A/D/A (analógico/digital/analógico). Esta placa interage com o sistema microcomputador através do barramento STD, ocupando um conector livre

do gabinete da máquina. Esta placa efetua três funções básicas: digitaliza o sinal analógico da voz; reconstrói a voz novamente a partir de voz digitalizada; e sintetiza a voz a partir dos parâmetros extraídos na análise LPC-10.

Para a primeira função, a placa conta com um conversor analógico-digital de 12 bits e filtro "anti-aliasing" apropriado. As amostras de voz quantizadas em 12 bits que representam a frase falada são imediatamente transferidas para o sistema que se encarrega de abrir um arquivo e armazená-las. A memória de massa onde são armazenados estes arquivos é um disco rígido (Winchester) com capacidade de 10 Mbytes. Com isto, não enfrentamos grandes problemas de limitação de área de armazenagem para os arquivos de coleta de voz, que são geralmente bastante extensos, se comparados ao seu par sintetizado. Apenas é preciso ter um mínimo de critério no uso da memória e apagar arquivos já fora de uso. A freqüência de amostragem é de 8 kHz.

Para reprodução de frases que foram apenas digitalizadas, a placa contém um conversor digital-analógico de 12 bits, seguido de filtro de reconstrução. Os arquivos que foram armazenados na memória de massa, compostos por amostras de 12 bits, podem ser redirecionados para reprodução novamente. Isto é importante para observação da gravação feita, bem como para comparar mais tarde com frases sintetizadas, realmente.

Para a execução da terceira função, a síntese da voz a partir de parâmetros LPC-10, a placa possui um Processador de Síntese de Voz da Texas Instruments, o TMS5220 [7]. Este é um componente dedicado, que implementa o filtro digital variante no tempo que simula o trato vocal, filtro este que é excitado ou pelo trem de impulsos ou pelo gerador de números aleatórios, dependendo do tipo de som ser sonoro ou surdo (ver **Fig.1**) A diferença com relação a esta figura é que o 5220 não realiza a pós-enfase, além de já conter o conversor D/A e um amplificador. Tornou-se então necessário implementar externamente ao 5220 a de-ênfase, analogicamente, porque o sinal entregue pelo 5220 já passou pelo D/A. Acrescentamos outro amplificador ao sinal de saída para termos um controle efetivo do volume da voz reproduzida. Este amplificador também é utilizado na reprodução da fala via D/A. Existe uma chave que possibilita direcionar uma das duas saídas para a reprodução.

Para a entrada da voz e a audição da mesma, utilizamos um monofone semelhante ao monofone de um telefone comum, mas com suas cápsulas trocadas por cápsulas dinâmicas. Toda entrada de sinal e posterior audição da voz, comprimida ou não, é feita via o monofone. A **Fig. 5** ilustra a placa A/D/A sob a forma de diagrama em blocos.

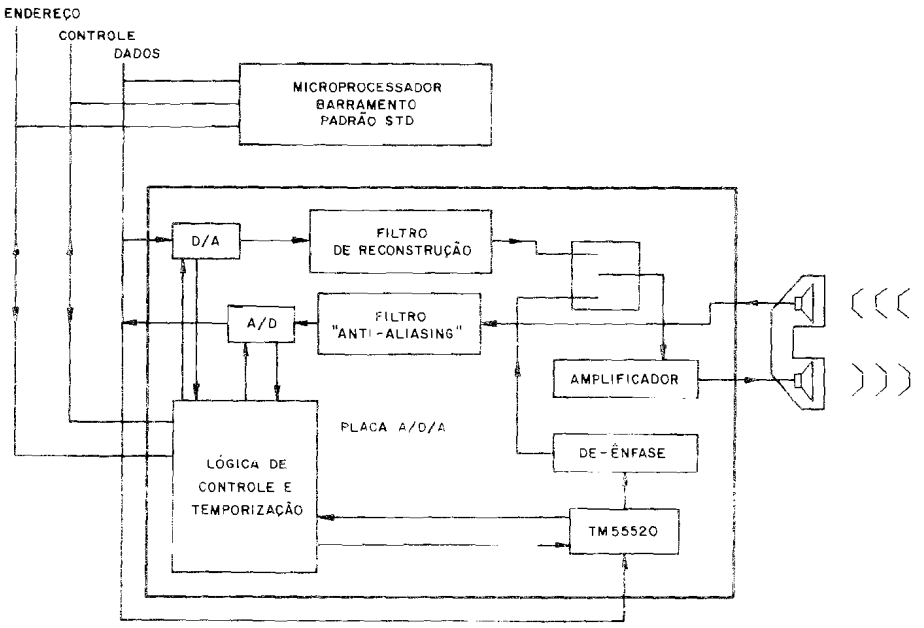

Figura 5. O sistema de simulação.

Em termos de "software", existem dois módulos básicos: a parte de controle do sistema, e o "software" encarregado efetivamente de promover a extração dos parâmetros na análise LPC.

O "software" de controle foi escrito em "assembler" do 8088 e sua função principal é de coordenação. É ele que abre os arquivos para a coleta das amostras de voz e que os direciona para a reprodução via D/A. Este "software" pode direcionar para a saída um arquivo composto por parâmetros LPC-10, que serão reproduzidos pelo circuito integrado dedicado 5220. O usuário interage com a máquina através deste módulo. O programa após carregado coloca na tela um menu que oferece as seguintes opções disponíveis: coleta de voz (C); reprodução via D/A (F); reprodução via 5220 (T); Ctrl C — retorna ao sistema operacional.

A partir de uma das três primeiras opções, o "software" pergunta o nome do arquivo. Na opção (C), o "software" automaticamente coloca a extensão .ABD ao nome escolhido para o arquivo a ser aberto. Ao usarmos a opção (F), o "software" pergunta o nome do arquivo .ABD a ser roteado para o D/A. No caso da opção (T), o arquivo que é gerado após a análise terá automaticamente a extensão .TMS acrescida ao nome originalmente escolhido, e será roteado para síntese no 5220.

Toda a parte de programação referente à análise da voz que resulta nos parâmetros LPC-10, ou seja, nos coeficientes de predição, no ganho e no valor do período fundamental, foi escrita em Turbo Pascal. Ao invocarmos o programa análise, ele pergunta qual o arquivo a ser analisado, que será um arquivo .ABD criado através da opção (C) do programa de controle. A análise consome bastante tempo de execução, porque são executados todos os algoritmos já descritos anteriormente. Este tempo gasto na análise é o limitador prático do tamanho dos arquivos de voz coletados, e não a memória de massa, o disco rígido. Se a frase gravada ultrapassar os 30 segundos de duração, a análise já começa a tomar horas de processamento. Ao fim, temos o arquivo .TMS, já formatado de acordo com as exigências do TMS5220, pronto para ser utilizado pela opção (T).

Como o Pascal é uma linguagem estruturada, de fácil compreensão, e como o programa é bem documentado, se torna bastante simples promover alterações nos algoritmos utilizados em todo o processo de análise. No início do programa, onde todas as variáveis, limiares, constantes, etc, são definidos, existem comentários indicando o que é cada um destes elementos, de acordo com os algoritmos. Alterar qualquer um deles para promover experimentos, se trata simplesmente de alterar o valor atribuído à variável, ou limiar etc, conforme o desejado.

Na verdade, sendo o Pascal uma linguagem estruturada, mesmo modificações de módulos de programa, ou seja, alterações nos algoritmos, são razoavelmente fáceis de serem efetuadas para quem possui alguma prática com a linguagem.

4. Resultados da Simulação

Como a percepção auditiva é bastante subjetiva, afirmar que uma mudança em alguma parte dos algoritmos realmente resultou em melhoria não é imediato. Para avaliar as mudanças introduzidas, elaboramos um teste de percepção que pode ser aplicado sempre que conveniente. O teste é de múltipla escolha, e o participante, após ouvir uma palavra, tem que escolher entre três opções escritas no papel à sua frente. Tanto a filosofia utilizada na construção como o exemplo da aplicação de um teste se encontram no Apêndice A.

O teste é composto de 50 quesitos, e o ouvinte tem 2 segundos para a sua tomada de decisão antes de ouvir o próximo conjunto de palavras. Como locutores, utilizamos duas vozes masculinas e duas vozes femininas intercaladas aleatoriamente, e inclusive com diferenças de sotaque. O auditório foi composto por 29 pessoas, de diferentes tipos de atividades e diferentes

níveis sociais. Repare que neste teste não há informação contextual. Apenas a informação existente em uma palavra isolada é usada na tomada de decisão pela resposta escolhida. Com isto, se torna mais fácil observar o tipo de som que o sistema está tendo dificuldade de tratar e reproduzir, para assim, se possível, aperfeiçoar o detalhe em questão.

O resultado da aplicação do teste no modelo análise/síntese como está atualmente alcançou 78% de acerto. O mesmo teste aplicado a outra platéia, mas sem nenhum processo de compressão, alcançou 95% de acerto. Considerando que muitas das partes não entendidas em uma conversação normal são deduzidas a partir do contexto da conversação, não é de se acreditar que problemas sérios de compreensão surjam em uma aplicação prática do sistema vocoder conforme descrito. No caso em questão, o teste tem como finalidade somente a avaliação da inteligibilidade do sinal reproduzido. Também utilizamos recursos de equipamentos de teste e medição para observação de resultados, mas só foi possível concluir sobre a real melhoria ou não de mudanças mais sutis nos algoritmos, a partir da aplicação do teste. De posse dos levantamentos feitos a partir dos resultados dos testes, observamos fatos bastante úteis para pesquisas futuras.

Considerando o universo de 50 palavras sintetizadas utilizadas no teste, a **Fig. 6** mostra um gráfico de percentual de acerto dentro do total de palavras que compõem o teste. A partir destes resultados, concluímos que ocorreram erros localizados, cujas causas devem ser estudadas com mais profundidade.

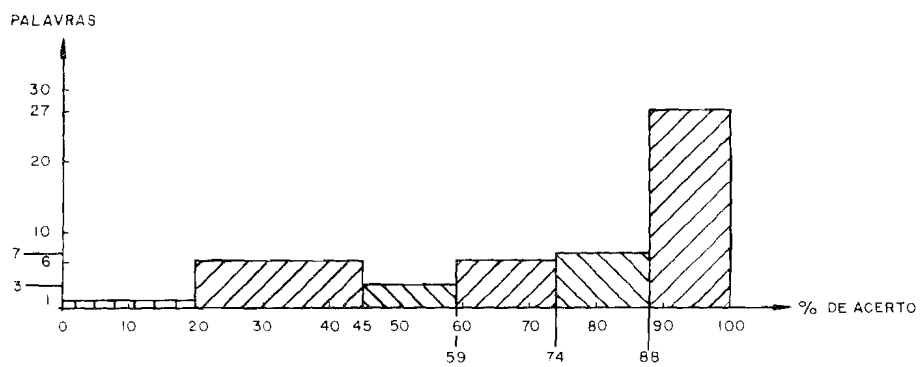

Figura 6. Distribuição percentual do acerto pelo total de palavras.

Conforme dito, utilizamos quatro vozes para nossas gravações, duas masculinas e duas femininas. A distribuição por acerto por locutor está na **Fig. 7**. Apesar da voz feminina ter mostrado melhor rendimento, podemos dizer que a máquina está bem ajustada, porque a diferença percentual entre os dois tipos de voz foi pequena.

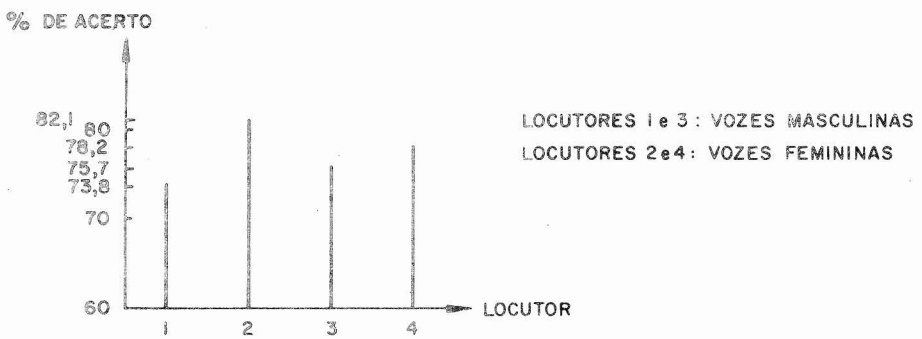

Figura 7. Distribuição percentual do acerto por locutor.

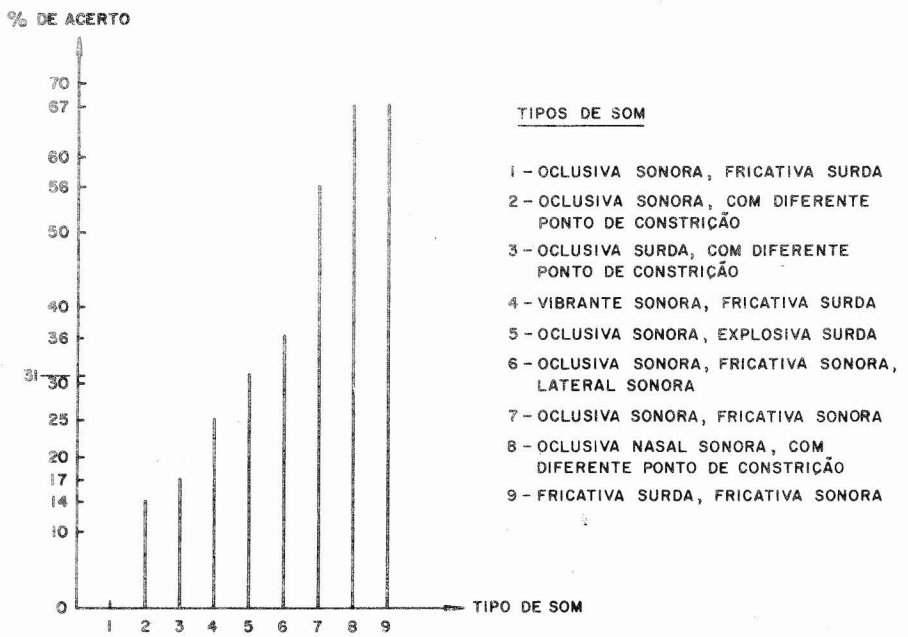

Figura 8. Distribuição percentual do acerto para alguns tipos de sons sintetizados.

A Fig. 8 mostra a distribuição percentual de acerto na identificação entre alguns tipos de sons sintetizados. A figura não cobre uma comparação entre todos os tipos de sons existentes, mas os que a princípio se mostraram mais problemáticos de serem discernidos. A observação do gráfico demonstra grande dificuldade de identificação entre sons oclusivos surdos (cato, tato) diferenciados apenas pelo ponto de constrição do fonema, entre sons oclusi-

vos sonoros (bica, dica) também diferenciados apenas pelo ponto de constrição do fonema, e em particular entre sons oclusivos sonoros e fricativos surdos (finco, som). Estudos mais profundos devem ser feitos para melhorar a reprodução e conseqüente identificação destes sons. Sons fricativos sonoros (zinco, juta) ou surdos foram bem diferenciados, bem como oclusivos nasais sonoros (neta, meta) diferenciados pelo ponto de constrição do fonema. Os demais encontraram notas razoáveis. Deve ser lembrado que estes percentuais são cálculos ajustados para os efeitos do contexto, conforme mencionado no Apêndice A.

Na verdade, para um resultado ainda mais consistente deste teste, seria necessário compor mais platéias, com elementos de atividades e níveis culturais os mais diferenciados possíveis, para obtenção de médias mais realistas. Observamos que elementos com melhor nível cultural, com maior vocabulário, obtêm um índice de acerto mais elevado. Da mesma forma, talvez fosse útil oferecer uma quantia em dinheiro pela participação no teste, para podermos de certa forma **nivelar** a boa vontade da platéia, porque observamos que o índice de acerto está relacionado também com o estado de espírito do ouvinte no momento da aplicação do teste. Por exemplo, o conjunto de palavras gingar, vingar, xingar, gerou 54% de acerto. Já o conjunto chá, já, vá, gerou 89% de acerto (cálculos feitos conforme indicado no Apêndice A). Ambas as questões são compostas por fricativos, diferindo entre som surdo ou sonoro. A diferença sensível no percentual de acerto deve ser imputada a razões subjetivas, como, por exemplo, o auditório utilizado não usar com freqüência palavras pertencentes ao primeiro conjunto.

O teste de avaliação é um ponto que merece estudos futuros mais profundos, pela sua vital importância para aprimoramento dos algoritmos.

5. Conclusões

Visando à construção de um vocoder, a simulação digital dos algoritmos envolvidos, apresentada neste trabalho, foi fundamental para assegurar o sucesso do empreendimento. Graças à experimentação de todas as condições esperadas, foi possível tirar conclusões e chegar a resultados de fundamental importância para chegarmos ao que consideramos uma boa qualidade de voz. Principalmente, a elaboração do algoritmo para decisão sonoro/surdo tirou grande proveito das facilidades de simulação. Inclusive, o equipamento final fará uso de um processador digital de sinais para efetuar a análise em tempo real, que trabalha em ponto fixo. Na simulação, foi também possível levar em consideração este detalhe, para avaliação de problemas de precisão numérica no ambiente futuro.

O teste de inteligibilidade demonstrou ser uma ferramenta de grande importância para a pesquisa de soluções otimizadas. Sua estruturação, modo de aplicação, formação de platéias, etc, deveriam ser material para pesquisas posteriores de pessoas com formação específica em fonética e fonologia, visando à elaboração de um teste mais efetivo e preciso. No momento, não é de nosso conhecimento nenhum estudo desse tipo feito no Brasil.

Os resultados finais obtidos foram bastante satisfatórios, conforme ficou evidenciado pelo percentual de acerto obtido na aplicação do teste de inteligibilidade, apesar deste não ser ideal e ter sido aplicado a poucas platéias.

Baseado no percentual de acerto obtido como resultado da aplicação do teste de inteligibilidade, e mantendo a taxa de transmissão baixa como a alcançada, não acreditamos que grandes melhorias possam ser realizadas tentando explorar o modelo LPC tradicional. Algum progresso deve ser possível aperfeiçoando ainda mais os métodos de extração de período fundamental e decisão sonoro/surdo, mas um novo patamar de qualidade só poderá ser alcançado como resultado da criação de um novo modelo, um nova forma de atacar o problema, pelo menos a taxas não excedendo os 2400 bit/s.

Apêndice A

Descrição do Teste de Inteligibilidade

Antes do advento das técnicas de codificação de voz, medir sua inteligibilidade era relativamente simples. Ruído, restrições na largura de faixa e limitações de potência eram as causas principais da degradação do sinal. Entretanto, o processamento digital de voz impõe suas formas de distorção próprias no sinal de voz. Em particular, técnicas que minimizam a largura requerida ao canal distorcem o sinal ainda mais. A complexidade destas distorções torna quase impossível predizer seus efeitos na inteligibilidade, a partir de medidas físicas feitas no sinal de voz processado. Assim é ainda necessário usar pessoas para este fim.

Na maior parte da interação verbal entre pessoas, existe mais de uma fonte de informação para se chegar ao conteúdo da mensagem: existe o estímulo da informação, isto é, a informação contida num segmento crítico da voz, e existe também muita informação dita contextual, ou informação contida nas circunstâncias da interação verbal. Estes fatores contextuais influenciam na percepção de fonemas, sílabas e outros elementos que compõem a conversação. O ambiente no qual ocorre a conversação (aviação, por exemplo) e a situação (aterrissagem, por exemplo) são informações que levam no seu contexto pistas para a compreensão da conversação em si.

Independentemente da importância do estímulo e contexto, a inteligibilidade é determinada pela habilidade do ouvinte em assimilar a informação destas fontes. Como a intuição sugere, a informação contextual é muito importante numa conversação, mas indivíduos diferem na sua capacidade de usar essa informação. Diferem mais do que na capacidade de usar outras características das quais a inteligibilidade depende. E este é um grande problema quando se pretende medir e controlar a inteligibilidade na comunicação falada, e portanto de voz processada por técnicas de processamento digital.

Uma maneira razoável de controlar o nível de estímulo contido na informação é controlá-lo através da razão sinal-ruído na comunicação. Mas controlar a informação contextual já não é tão fácil. Informação contextual pode ser melhor definida como a totalidade de fatores, verbais e extra-verbais, que influenciam na probabilidade da ocorrência de um evento elementar na fala [8], [9], e daí a grande dificuldade de exercer um controle efetivo sobre eventos tão aleatórios. Em um teste, qual a quantidade de informação contextual que deveria ser dada em um dado texto? O ideal seria dar a mesma quantidade que o ambiente, ou o meio no qual o equipamento será usado dará. Mas isto é praticamente impossível de se saber, mais ainda que um determinado equipamento pode se usado em situações as mais diversas. Assim, apesar de não reproduzir a condição real, a forma mais legítima de avaliar inteligibilidade em um sistema é usar modelos sem informação contextual. Mas mesmo isto é difícil, porque os ouvintes, ou pior ainda; apenas alguns dos ouvintes, sempre conseguem reduzir a incerteza do teste, através de associações entre palavras, diferenças no vocabulário mais usual entre os componentes da platéia e outros motivos.

Testes de múltipla escolha se prestam bem para reduzir a quantidade de informação contextual em um teste. Outro problema é a escolha do conteúdo do teste e como este será organizado. O processo de escolha de palavras e sua distribuição dentro do teste é essencialmente o mesmo que é usado por educadores para eliminar (ou reduzir) o efeito da escolha aleatória da resposta nos resultados de testes de múltipla escolha. Para pelo menos minimizar este efeito, a incerteza das deduções a priori tem que ser maximizada, ou seja, a contribuição de fatores contextuais tem que se reduzida. A fórmula que se segue é utilizada para dar resultados de testes considerando este problema, o que conduziria então a resultados mais reais. Supondo que todos os ítens são respondidos, tem-se de [10] que

$$P_C = \frac{P_r - \frac{P_w}{n-1}}{T} \qquad \text{(A.1)}$$

onde P_C é a percentagem de acerto, ajustada para os efeitos do contexto, P_r é o número de respostas corretas, P_w é o número de respostas erradas,

n é o número de escolhas em cada ítem e T é o número total de ítens do teste.

Esta fórmula permite, por exemplo, uma comparação significativa entre resultados de testes com tamanhos diferentes. Entretanto, a validade do resultado parte de duas premissas. A primeira é que os conjuntos de opções de respostas são compostos por unidades igualmente atraentes. Por exemplo, um conjunto de respostas como gato, pato, caco, abacaxi afetaria a validade do resultado. A segunda premissa é que a resposta do ouvinte a um quesito depende de uma única decisão discriminativa. Quando o estímulo da incerteza é confinado a um único fonema ou detalhe para a correta compreensão, a fórmula funciona bem.

Quando da apresentação do teste à platéia, outra pergunta que surge é quanto ao tempo indicado para os ouvintes tomarem a decisão da resposta nos ítens do teste de múltipla escolha. Em testes deste tipo, feitos para a língua inglesa, o tempo ótimo para responder cada ítem está entre 1,5 e 4 segundos, aproximadamente.

A variação dos tipos de palavras no universo que compõe o teste também é fator a considerar. Se obtemos um certo resultado para um teste e depois aumentamos a variedade de tipos de sons cobertos pelo teste, sem melhorarmos a qualidade do som reproduzido, é previsível uma queda no percentual de acertos.

É recomendável no teste de equipamentos que reproduzam a voz de alguma forma, o uso de mais de um locutor durante as gravações. O ideal seria dispor de tantos locutores quantos são os ouvintes. Quanto ao número de ouvintes, experiências anteriores indicam que 8 a 10 ouvintes constituem uma platéia razoável para um dado teste, particularmente se são pessoas conscientes e com boa vontade, para que a consistência nos resultados possa ser mantida. Um pouco mais complicado é a escolha do número de opções de respostas dadas ao ouvinte. O DRT (Diagnostic Rhyme Test), por exemplo, permite a escolha entre duas opções, e o MRT (Modified Rhyme Test) entre seis [10].

Além do número de respostas, a escolha apropriada de opções deve considerar as características fonéticas a serem analisadas. Com isto, um estudo dos resultados colhidos evidenciará qual característica sonora os algoritmos estão tendo dificuldades de reproduzir. As características principais que consideramos são as seguintes: sons sonoros, sons surdos, sons nasais, sons orais, sons fricativos, sons oclusivos, sons laterais e sons vibrantes. No nosso teste, escolhemos somente palavras com estas características, que são aquelas observadas de modo mais marcante na nossa língua.

Também devem ser considerados a natureza e o volume do ruído de fundo durante a aplicação do teste. Assumimos o ruído ambiente normal, como ar condicionado, telefone, etc, como o ruído de fundo presente no recinto de aplicação do teste.

O teste que propomos elimina ao máximo a informação contextual, usando palavras fora de frases e o método de múltipla escolha. O tempo dado para que seja escolhida a resposta a cada ítem é de 2 segundos, e são três as opções de resposta para cada um deles. O teste é composto por 50 perguntas.

A seguir, a título de ilustração, apresentamos um teste completo, na forma em que é apresentado à platéia. O campo **Teste Aplicado** se refere ao tipo de teste que estamos aplicando: se voz simplesmente reproduzida ou digitalizada e comprimida.

Teste de Identificação de Fonemas

Ouvinte:
Data:
Teste Aplicado:

1)	paca	vaca	faca
2)	mico	bico	pico
3)	gato	tato	cato
4)	seus	deus	zeus
5)	grama	trama	drama
6)	gingar	vingar	xingar
7)	feio	veio	meio
8)	vanda	panda	banda
9)	gosta	posta	costa
10)	finco	cinco	zinco
11)	doca	foca	toca
12)	chá	já	vá
13)	vala	mala	bala
14)	fada	nada	dada
15)	beta	meta	teta
16)	disso	viço	nisso
17)	marca	barca	parca
18)	dona	tona	nona
19)	vingo	bingo	xingo
20)	parda	sarda	farda
21)	vê-lo	sêlo	pêlo
22)	adaga	apaga	afaga
23)	adora	arvora	arbora
24)	grife	gripe	grite

25)	whisky	disque	risque
26)	tampa	tanga	tanta
27)	bica	dica	mica
28)	canta	cana	cama
29)	luva	uiva	ruiva
30)	maca	mapa	mata
31)	som	dom	bom
32)	seta	neta	meta
33)	garra	jarra	barra
34)	dumbo	jumbo	bumbo
35)	luta	juta	guta
36)	jurar	durar	furar
37)	muito	mundo	minto
38)	caco	calo	cato
39)	risco	disco	fisco
40)	agora	afora	adora
41)	cua	tua	pua
42)	juno	junho	julho
43)	caio	caro	caco
44)	terra	berra	guerra
45)	afro	abro	agro
46)	alar	assar	achar
47)	com	cal	cão
48)	mal	má	mão
49)	sim	sem	se
50)	cinto	cio	cito

Referências

[1] "STDMG STD BUS Specification and Practice", Pro-Log Company, Documento no.10689E, Outubro 1984.

[2] L. R. Rabiner e R. W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, 1978.

[3] B. Gold e L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain", The Journal of the Acoustical Society of America, vol. 46, Agosto 1969, pp. 442-448.

[4] F. A. G. Ferreira, "Implementação de um Algoritmo para Extração de "Pitch" em Sinais de Voz", Tese de Mestrado, Universidade de Brasília, Departamento de Engenharia Elétrica, Dezembro 1985.

[5] J. D. Markel e A. H. Gray, "A Linear Prediction Vocoder Simulation Based Upon the Auto-Correlation Method", IEEE Transactions on Acoustics and Speech Signal Processing, vol. ASSP-22, Abril 1974, pp.124-134.

[6] B. Gold, "Note on Buzz-Hiss Detection", The Journal of the Acoustical Society of America, vol. 36, no. 9, Setembro 1964, pp.1659-1661.

[7] "TMS5220 Voice Synthesis Processor Data Manual", Texas Instruments, Junho 1981.

[8] G. Fairbanks, "Test of Phonemic Differentiation: The Rhyme Test", The Journal of the Acoustical Society of America, vol. 30, 1958, pp.596-600.

[9] G. A. Miller, G. A. Heise e W. Lichten, "The Intelligibility of Speech as a Function of the Context of the Test Materials", Journal of Experimental Psychology, vol. 41, 1951, pp.329-335.

[10] W. D. Voiers, "Evaluating Processed Speech Using the Diagnostic Rhyme Test", Speech Technology, Janeiro/Fevereiro 1983.

MARCO ANTONIO T. PEREIRA concluiu o curso de Engenharia Elétrica na Universidade de Brasília em junho de 1979 e obteve o grau de Mestre em Engenharia Elétrica na Syracuse University em setembro de 1983. Entre 1979 e 1981 e entre 1983 e 1985, trabalhou na equipe de desenvolvimento da Coencisa Indústria de Comunicações S.A. Atualmente, é sócio-gerente da Kernel Informática Ltda, Brasília,DF. Suas áreas de interesse são processamento digital de sinais, comunicação de dados, arquitetura de microprocessadores.

FERNANDO ACATAUASSU G. FERREIRA concluiu o curso de Engenharia Elétrica na Universidade de Brasília (UnB) em dezembro de 1977. Cursou pós-graduação em Eletrônica Digital na PUC-Rio e especialização em Controle de Processos na UnB. Obteve o grau de Mestre em Engenharia Elétrica pela UnB em dezembro de 1985. Entre 1978 e 1985 trabalhou na equipe de desenvolvimento da Coencisa Indústria de Comunicações S.A. Atualmente é sócio-gerente da Kernel Informática Ltda, Brasília, DF. Suas áreas de interesse são processamento digital de sinais, comunicações de dados, redes de comunicações.