UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ELETRÔNICA E DE COMPUTAÇÃO

ALGORITMOS EFICIENTES DE EXTENSÃO EM FREQÜÊNCIA DO SINAL DE TELEFONE

Autora: Cássia Valentini Botinhão

Orientadora: Mariane Rembold Petraglia

Examinador: Luiz Pereira Caloba

Examinador: José Gabriel R. C. Gomes

DEL

Agosto de 2006

AGRADECIMENTOS

Gostaria de agradecer a Deus por todos os sonhos, oportunidades e metas que tenho alcançado.

Aos meus pais, João Aldo e Maria Nina, e minha irmã, Carine, pelo amor, amizade, apoio e paciência que sempre me dedicaram.

Aos amigos que me ajudaram durante esse período de graduação.

Por fim, gostaria de agradecer a minha professora e orientadora Mariane Rembold Petraglia, pelos ensinamentos passados e pelo apoio dado a mim durante os cinco anos de curso e no desenvolvimento deste trabalho.

RESUMO

Cássia Valentini Botinhão
UFRJ - Escola Politécnica

Projeto de Final de Curso
Agosto de 2006

Palavras-chave: extensão em freqüência, processamento digital de sinais de voz, banco de filtros, filtragem adaptativa, redes neurais.

Algoritmos Eficientes de Extensão em Freqüência do Sinal de Telefone

Esse trabalho trata do problema de extensão da banda de freqüência de sinais de voz no telefone realizada com o intuito de melhorar a percepção auditiva desses sinais. A idéia básica é desenvolver um algoritmo que seja capaz de, a partir da informação contida no sinal de banda estreita transmitido atualmente nos sistemas de telefonia, gerar artificialmente a banda que foi perdida por causa do limite imposto pela taxa de transmissão do sistema. São propostos dois algoritmos que realizam a extensão explorando de maneiras distintas as relações existentes entre as componentes freqüênciais perdidas e as que continuam presentes no sinal de telefone recebido pelo assinante. Os resultados obtidos para ambos algoritmos foram analisados tanto quantitativamente quanto subjetivamente, através de testes realizados com voluntários que avaliaram os resultados em relação à extensão proporcionada pelo algoritmo e à presença ou não de ruídos e artefatos no sinal estendido.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O compromisso entre a qualidade de transmissão e a taxa de amostragem utilizada atualmente na linha telefônica definiu uma taxa de 8000 Hz para a transmissão. De acordo com o teorema da amostragem de Nyquist, para que um sinal seja transmitido sem perdas que comprometam a sua reconstrução, sua banda de freqüência, ou seja, a maior freqüência contida no sinal, deve ser menor que a metade da taxa de amostragem que o sinal está sendo submetido. O sinal de voz, que ocupa uma faixa de 100 Hz a 8000 Hz é então filtrado para ocupar uma banda de até 4 kHz para manter a integridade do sinal amostrado a 8000 Hz, conforme ilustrado na Figura 1.1.

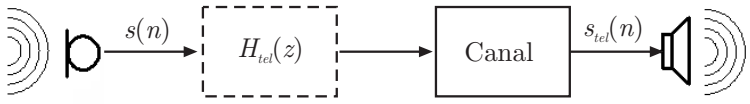


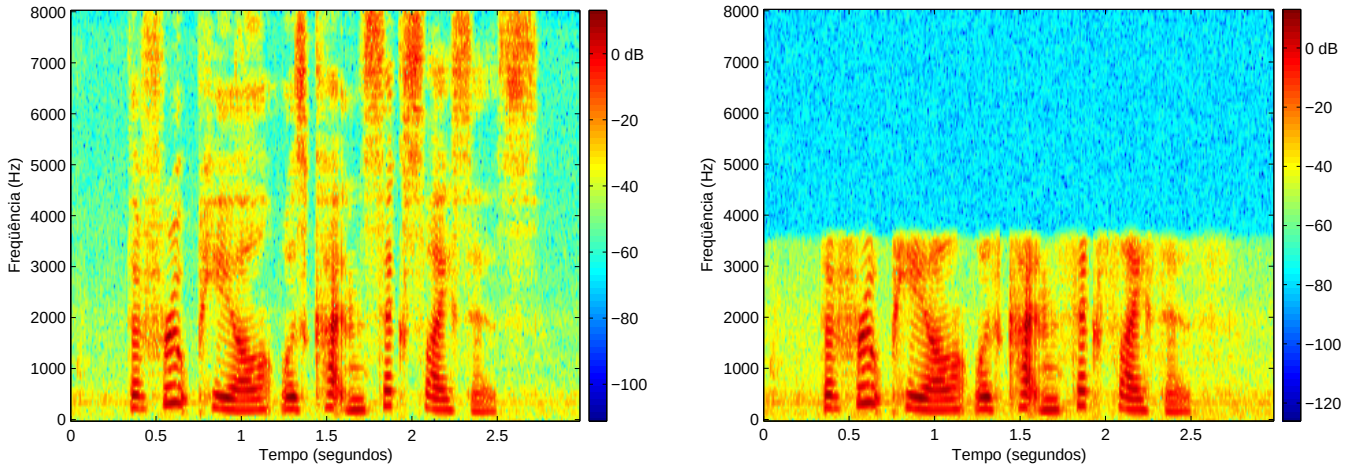
Figura 1.1: Estrutura do sistema atual de telefonia

A Tabela 1.1 descreve a atenuação que o filtro $H_{tel}(z)$, caracterizado pela *International Telecommunication Union - Telecommunication Standardization Sector* (ITU-T), impõe no sinal $s(n)$ a ser transmitido na linha telefônica [1]. Nesse processo de filtragem as componentes de freqüência de 4000 Hz a 8000 Hz são consideradas perdidas e as componentes do intervalo de 3400 Hz a 4000 Hz e de 100 a 300 Hz, correspondentes bandas de transição do filtro $H_{tel}(z)$, se encontram atenuadas no sinal de banda estreita $s_{tel}(n)$.

A Figura 1.2(a) mostra o espectrograma do sinal de voz a ser transmitido pela linha telefônica que de acordo com a Figura 1.1 corresponde ao sinal $s(n)$. O espectrograma

Freqüência (Hz)	Atenuação(db)
100 - 300	0 - 10
300 - 3400	0
3400 - 4000	0 - 18

Tabela 1.1: Caracterização da resposta em freqüência do filtro de canal ITU-T G.712



(a) Sinal de banda larga

(b) Sinal de banda estreita

Figura 1.2: Espectrograma dos sinais envolvidos na transmissão

disponibiliza a relação existente entre a potência do sinal no tempo e na freqüência. Podemos ver que o sinal de banda larga $s(n)$ contém componentes de freqüência com alta potência na região de rejeição dos filtros utilizados no atual sistema de telefonia. Essas componentes já não são encontradas na Figura 1.2(b), onde é possível visualizar o espectrograma do sinal de banda estreita $s_{tel}(n)$ que chega ao assinante no aparelho de telefone.

As componentes perdidas comprometem a qualidade e a inteligibilidade do sinal de voz. Se no começo dos sistemas de telefonia essas perdas não representavam grande insatisfação por parte dos assinantes, atualmente a qualidade do sinal de telefone, se comparada à de outras fontes de áudio como rádio e o CD (*Compact Disk*), é uma queixa mais freqüente.

Segundo [1], sinais de voz limitados em freqüência possuem inteligibilidade de 99% para uma sentença completa, e para sílabas esse valor é de 90%. Em uma análise mais detalhada das características do sinal de voz, podemos prever que a perda de componentes de altas freqüências deve proporcionar maior impacto na inteligibilidade dos fonemas consonantais, particularmente para os fonemas fricativos tais como o /s/, /sh/, /ch/, /x/, /th/, pois seus espectros se estendem significativamente acima do intervalo de banda utilizado na transmissão.

Para implementar comunicação com sinais de banda larga seria necessário modificar os *links* de transmissão, o que demandaria um grande investimento. Uma solução alternativa é o desenvolvimento de algoritmos de extensão artificial da banda de freqüência dos sinais que chegam no receptor, como ilustrado na Figura 1.3.

Os algoritmos de extensão de banda são realizados com o intuito de gerar as componentes perdidas no sinal de banda estreita a partir somente da informação contida nesse sinal. Parte-se do pressuposto que a banda estreita tem consigo informações capazes de

reconstruir a um certo nível a banda perdida, de acordo com a teoria da informação [2].

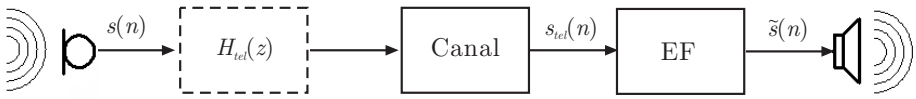


Figura 1.3: Extensão em Freqüência (EF) no contexto da transmissão de um sinal telefônico

Parte significativa dos algoritmos de extensão em freqüência apresentados na literatura utiliza a estrutura contida no modelo do trato vocal humano para realizar a extensão.

Este modelo tenta simular as modificações que ocorrem na corrente de ar que chega à cavidade bucal e nasal, ilustradas na Figura 1.4(a), através do filtro $1/A(z)$, no qual $A(z)$ se refere ao filtro de erro de predição linear (LPC). O sinal $e(n)$ que entra no filtro $1/A(z)$ representa, na formação da voz humana, a corrente de ar que chega a essas cavidades, e é denominado de sinal de excitação. A produção da fala humana é caracterizada, portanto, pelos coeficientes $a(n)$ do filtro de erro de predição $A(z)$, que são calculados a partir do sinal de voz $s(n)$, e pelo sinal $e(n)$ que serve de excitação para o inverso desse filtro. Como o sinal de excitação tem características de um ruído branco, a resposta em freqüência de $1/A(z)$ é considerada o envelope espectral do sinal $s(n)$.

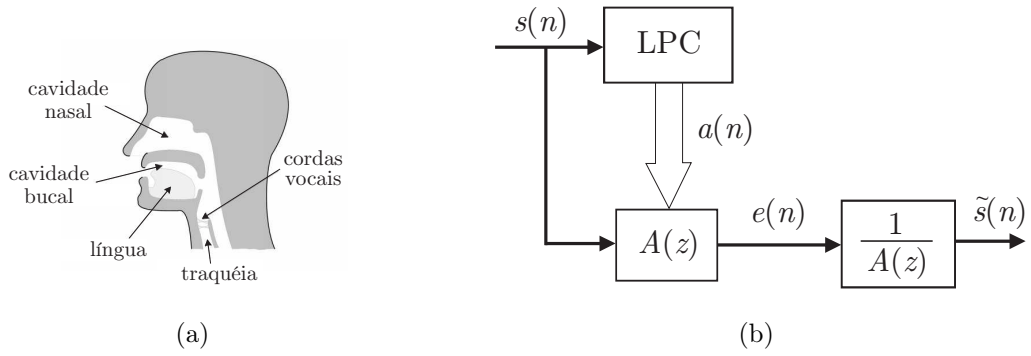


Figura 1.4: (a) Trato vocal; (b) modelo da formação da voz humana

No problema de extensão de banda de freqüência é necessário estimar os coeficientes $a_{BL}(n)$ do modelo de síntese de voz para o sinal de banda larga, e o sinal de excitação $e_{BL}(n)$ para gerar o sinal de banda larga, de acordo com a Figura 1.5.

As técnicas mais utilizadas para gerar o sinal de excitação de banda larga $e_{BL}(n)$ a partir do sinal $e_{BE}(n)$ de banda estreita são:

- modular $e_{BE}(n)$ com uma função cosseno cuja freqüência é constante [1] ou varia de acordo com o *pitch* do sinal processado [3];

- realizar operações não lineares no sinal $e_{BE}(n)$, como, por exemplo, elevá-lo ao quadrado ou ao cubo [4], a fim de acrescentar harmônicos ao sinal.

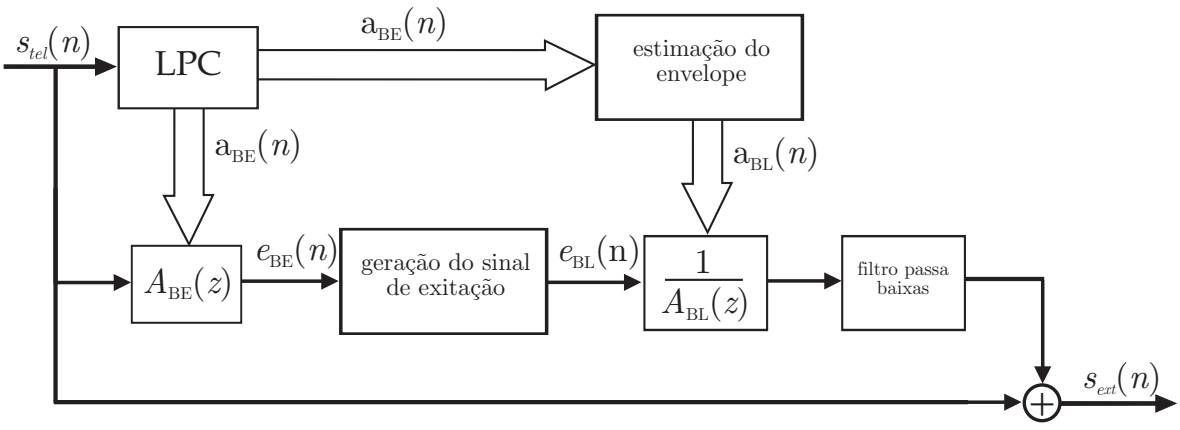


Figura 1.5: Estrutura mais comum de algoritmos de extensão em freqüência

A estimação dos coeficientes $a_{BL}(n)$ a partir de $a_{BE}(n)$, realizada para obtenção do filtro do envelope espectral $1/A_{BL}(z)$, é tradicionalmente implementada através das seguintes ferramentas:

- mapeamento linear, que consiste em encontrar uma matriz que aplicada ao vetor formado pelos coeficientes $a_{BE}(n)$ gere uma estimativa dos coeficientes $a_{BL}(n)$. Pode-se também calcular uma matriz que relacione grandezas, como *line spectral frequencies* (LSF), do sinal de banda larga e do sinal de banda estreita [5];

- rede neural, cujo vetor de entrada é $a_{BL}(n)$ e cujo par de entrada e saída de treinamento é composto pelos vetores $a_{BL}(n)$ e $a_{BE}(n)$ respectivamente. A rede neural é capaz de realizar um mapeamento não linear entre esses coeficientes;

- *codebook*, um banco de dados contendo pares de $a_{BL}(n)$ e $a_{BE}(n)$ construído a partir de um banco de vozes. O mapeamento é então realizado acessando esse banco de dados e procurando o par correspondente através de métricas.

Segundo [4], os algoritmos de extensão em freqüência que utilizam *codebook* apresentam melhores resultados em métricas tanto quantitativas quanto subjetivas. Porém, a complexidade computacional destes algoritmos é muito grande.

Neste trabalho foram desenvolvidos dois novos algoritmos para extensão em freqüência, que apresentam baixa complexidade computacional.

O primeiro algoritmo desenvolvido utiliza filtros lineares para gerar os sinais das bandas de mais altas freqüências a partir de componentes de baixas freqüências. Esses filtros são encontrados para duas classes de voz, fonemas vozeados e não vozeados, e uma rede neural classificatória determina a qual classe cada trecho do sinal de voz processado pertence. O segundo algoritmo utiliza um estimador não linear (rede neural) para obter os coeficientes $a_{BL}(n)$ capazes de gerar o sinal estendido. Ambos algoritmos implementam a extensão ao nível das sub-bandas de freqüência do sinal de banda estreita.

Capítulo 2

Fundamentos Teóricos

Neste capítulo será apresentada a teoria que possibilitou a elaboração dos algoritmos de extensão em freqüência. Essa teoria engloba tópicos de predição linear, como o filtro de erro de predição, de processamento multitaxas, tal como banco de filtros modulado por cosseno, de filtragem adaptativa, como o método dos mínimos quadrados, e conceitos de treinamento e teste de redes neurais.

2.1 Modelo

2.1.1 Predição linear

O conceito de predição linear envolve estimar o valor de um processo $x(n)$ no instante (n), a partir somente das amostras $\{x(n-1), x(n-2), ..., x(n-M)\}$, sendo M a ordem do preditor, como descrito em [6]:

$$\hat{x}(n|X_{n-1,n-M}) = \sum_{k=1}^{M} h_0(k)x(n-k) \tag{2.1}$$

sendo $X_{n-1,n-M} = [\ x(n-1)\ \ x(n-2)\ \ ...\ \ x(n-M)\ ]^T$.

O filtro de predição linear de ordem M tem a seguinte função de transferência:

$$H_{0_M}(z) = \sum_{k=0}^{M-1} h_0(k+1)z^{-k} \tag{2.2}$$

A definição de erro de predição segue a seguinte equação:

$$e(n) = x(n) - \hat{x}(n|X_{n-1,n-M}) \tag{2.3}$$

Relacionando as Equações (2.1) e (2.3) é possível verificar que os coeficientes do filtro de erro de predição são descritos da seguinte maneira:

$$a_M(m) = \begin{cases} 1, & m = 0 \\ -h_0(m), & m = 1, 2, ..., M \end{cases} \tag{2.4}$$

De acordo com essas definições é obtida uma relação entre as funções de transferência do filtro de predição linear $H_{0_M}(z)$ e do filtro de erro do predição linear $A_M(z)$, ilustrada na Figura 2.1:

$$\begin{aligned} A_M(z) &= \sum_{k=0}^{M} a_M(k) z^{-k} && (2.5) \\ &= 1 - \sum_{k=1}^{M} h_0(k) z^{-k} \\ &= 1 - H_{0_M}(z) z^{-1} && (2.6) \end{aligned}$$

Pode-se relacionar de acordo com [6], algumas propriedades do filtro do erro de predição que nos serão úteis para introduzir o modelo do trato vocal humano:

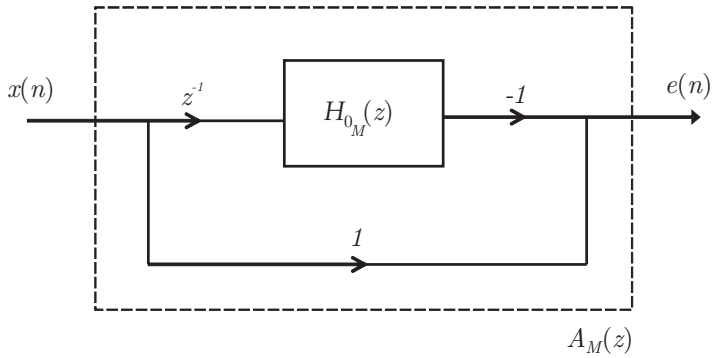


Figura 2.1: Filtros de predição linear $H_{0_M}(z)$ e de erro de predição linear $A_M(z)$

- $A_M(z)$ tem fase mínima, ou seja, todos os seus zeros estão dentro ou sobre o círculo unitário;
- Para M grande, o filtro do erro de predição $A_M(z)$ branqueia o processo $x(n)$, ou seja, retira informação redundante a cerca do processo.

2.1.2 Processo auto-regressivo

Um processo é dito um processo auto-regressivo (AR) de ordem M quando seu valor atual $x(n)$ é uma combinação linear de M valores anteriores desse mesmo processo mais uma inovação, ou seja, um ruído branco $v(n)$:

$$x(n) = v(n) - a_1 x(n-1) ... - a_M x(n-M) \tag{2.7}$$

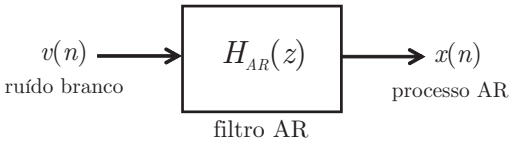


Figura 2.2: Filtro $H_{AR}(z)$

Quando $x(n)$ é um processo auto-regressivo (AR) de ordem M, o erro $e(n)$ de um filtro de predição linear $H_0(z)$ associado a esse processo será um ruído branco. De acordo com a definição de processo AR é possível relacioná-lo com o filtro de predição linear da seguinte maneira:

$$H_{AR}(z) = \frac{1}{A_M(z)} \tag{2.8}$$

no qual $H_{AR}(z)$ está ilustrado na Figura 2.2.

Essa representação de processos estacionários será explorada no modelo que aproxima a formação da fala humana.

2.1.3 Formação da fala humana

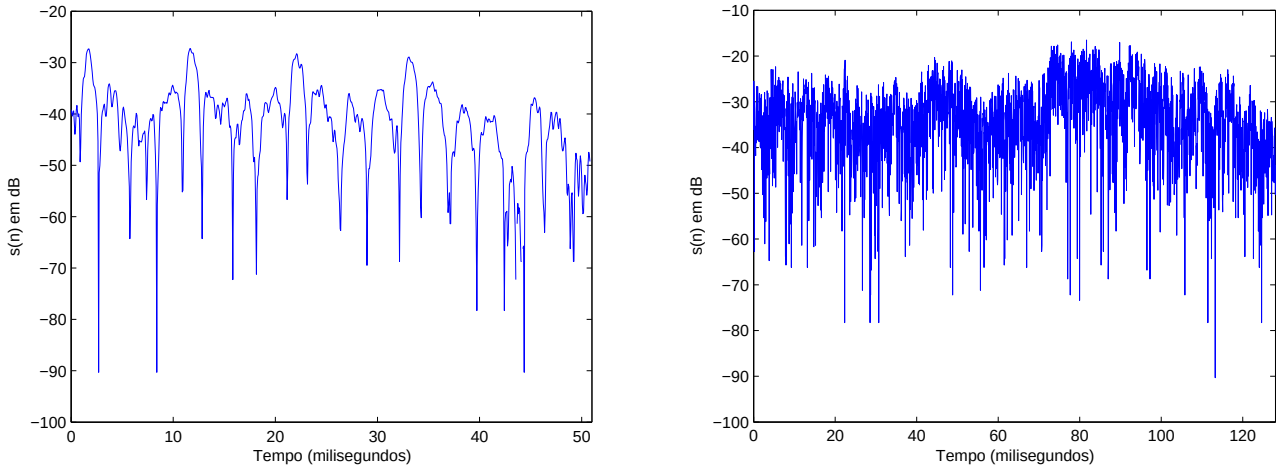
Para produzir a voz humana a língua bombeia ar da traquéia para a cavidade nasal e bucal, ilustradas na Figura 1.4(a). A corrente de ar bombeada segue pelo chamado trato vocal, que se estende desde a abertura das cordas vocais até a boca seguindo uma parte pela cavidade nasal.

O que diferencia a formação de um fonema para outro está principalmente nas características de ressonância e de reflexão do trato vocal e na geração do sinal de excitação do trato vocal. As reflexões e ressonâncias que esse sinal sofre no trato vocal dependem dos formantes, que são freqüências que podem ser alteradas, por exemplo, pelo movimento da língua.

O sinal de excitação é gerado pelo movimento das cordas vocais produzindo dois tipos distintos de fonemas:

- Fonemas vozeados: são gerados quando as cordas vocais estão vibrando, abrindo e fechando. O sinal gerado, portanto, é periódico com período fundamental chamado de período de *pitch*, como ilustrado na Figura 2.3(a). São classificados como vozeados, neste trabalho, as vogais e os fonemas nasais.

- Fonemas não vozeados: são gerados quando as cordas vocais estão abertas, o que resulta em um sinal com características similares a um ruído, como mostrado na Figura 2.3(b). Estão nessa classe a maioria dos fonemas fricativos e plosivos (/t/,/p/,/th/ dentre outros). É conhecido que os fonemas fricativos contém ruídos [7].



(a) Fonema vozeado /n/

(b) Fonema não vozeado /sh/

Figura 2.3: Sinais de voz no tempo

2.1.4 Modelo da formação da fala humana

O modelo da formação da fala humana deve ser capaz de simular o sinal de excitação $e(n)$ e o trato vocal $H_{TV}(z)$, presentes na Figura 2.4, visto que são estes os responsáveis pela caracterização de um fonema.

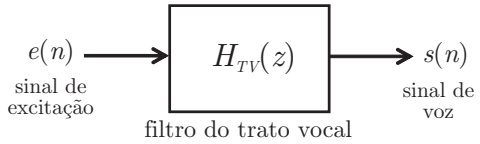
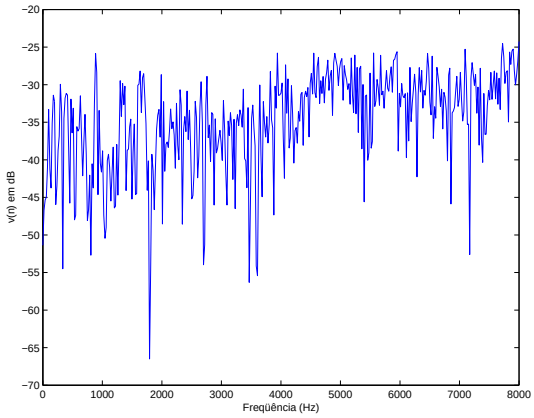


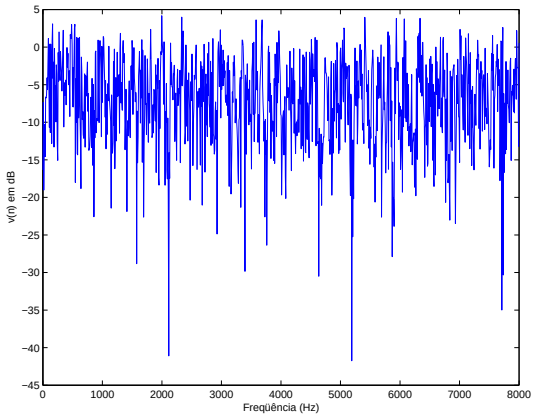
Figura 2.4: Estrutura do modelo com filtro do trato vocal $H_{TV}(z)$ e sinal de excitação $e(n)$

O sinal de excitação tem a propriedade de ter seu espectro constante para todas as freqüências, como é ilustrado nas Figuras 2.5. Portanto, podemos considerar que a resposta em freqüência de um filtro que modela o trato vocal humano é o envelope espectral do sinal de voz $s(n)$, como pode ser visto nas Figuras 2.6. Isso é válido em um curto período de tempo (de 10ms a 30ms) no qual o sinal de voz pode ser considerado como sendo estacionário, já que as características do trato vocal e o tipo de sinal de excitação mudam relativamente devagar.

O filtro $H_{TV}(z)$ que modela o trato vocal, representando o envelope espectral, pode ser descrito como um filtro somente pólos de baixa ordem com a seguinte resposta em freqüência:

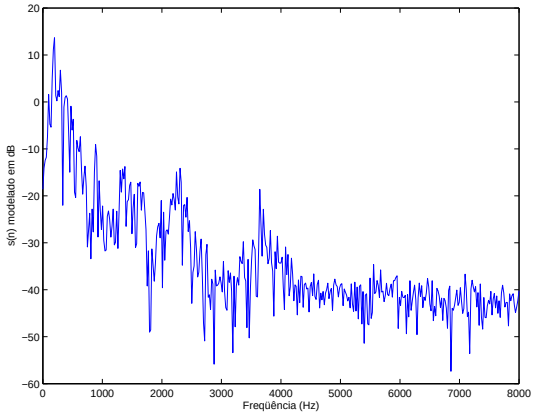


(a) Fonema vozeado /n/

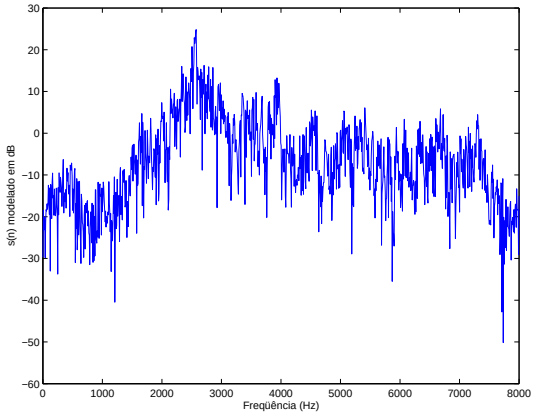


(b) Fonema não vozeado /sh/

Figura 2.5: Espectros do sinal de excitação



(a) Fonema vozeado /n/



(b) Fonema não vozeado /sh/

Figura 2.6: Espectros do sinal de voz modelado

$$H_{TV}(e^{j\omega}) = \frac{\sigma}{A(e^{j\omega})} \tag{2.9}$$

$$= \frac{\sigma}{1 - \sum_{i=1}^{p} a_i e^{j\omega i}} \tag{2.10}$$

Comparando as Equações (2.6), (2.8) e (2.10) pode-se relacionar o filtro do erro de predição linear, $A_M(z)$, o filtro do modelo AR, $H_{AR}(z)$, e o filtro que modela o trato vocal, $H_{TV}(z)$, pela seguinte expressão:

$$H_{TV}(z) = H_{AR}(z) = \frac{1}{A_M(z)} \tag{2.11}$$

A Figura 2.7 ilustra a estrutura que tenta reconstruir os dois tipos de fonemas citados na Subseção 2.1.3. Um gerador de ruído branco e um gerador de sinal periódico no tempo, cujo período é o inverso da freqüência de *pitch*, são utilizados como fontes para gerar o

sinal de excitação de fonemas não vozeados e vozeados, respectivamente.

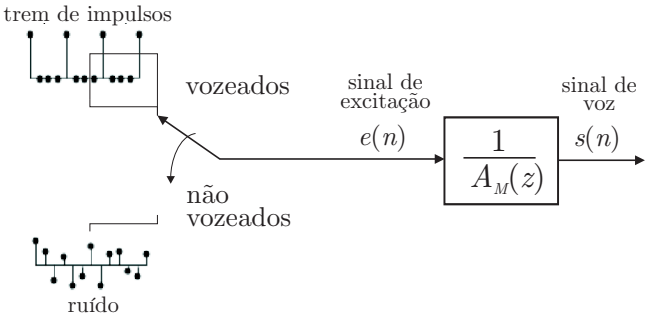


Figura 2.7: Modelo do filtro da fonte vocal para o processo de formação da fala humana

Dessa maneira para simular o processo de geração do sinal de voz é preciso identificá-lo como vozeado ou não vozeado (no primeiro caso é necessário também encontrar a freqüência de *pitch*) e calcular os coeficientes do filtro que modela o trato vocal, ou seja, encontrar os coeficientes do modelo auto-regressivo de ordem p do sinal.

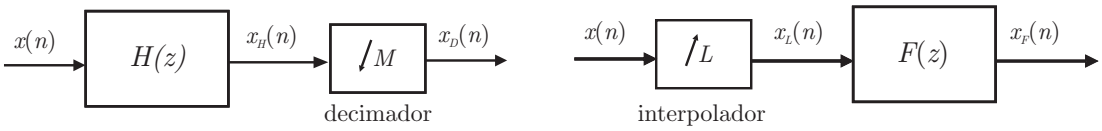
Uma maneira de se encontrar os coeficientes $\{a_1, a_2, ...a_p\}$ do filtro H_{AR} é utilizar o método da covariância modificada [8]. Esse método minimiza os erros de predição de acordo com o critério dos mínimos quadrados.

No contexto de extensão em freqüência, o modelo do trato vocal será utilizado como uma ferramenta que identifica os parâmetros que devem ser estimados a partir do sinal de banda estreita para gerar o sinal de banda larga. Dessa maneira, como mencionado no Capítulo 1, o algoritmo de extensão em freqüência irá estimar o envelope espectral, ou seja, H_{AR}, e o sinal de excitação $e(n)$, que serão utilizados para reconstruir o sinal.

2.2 Processamento Multitaxas

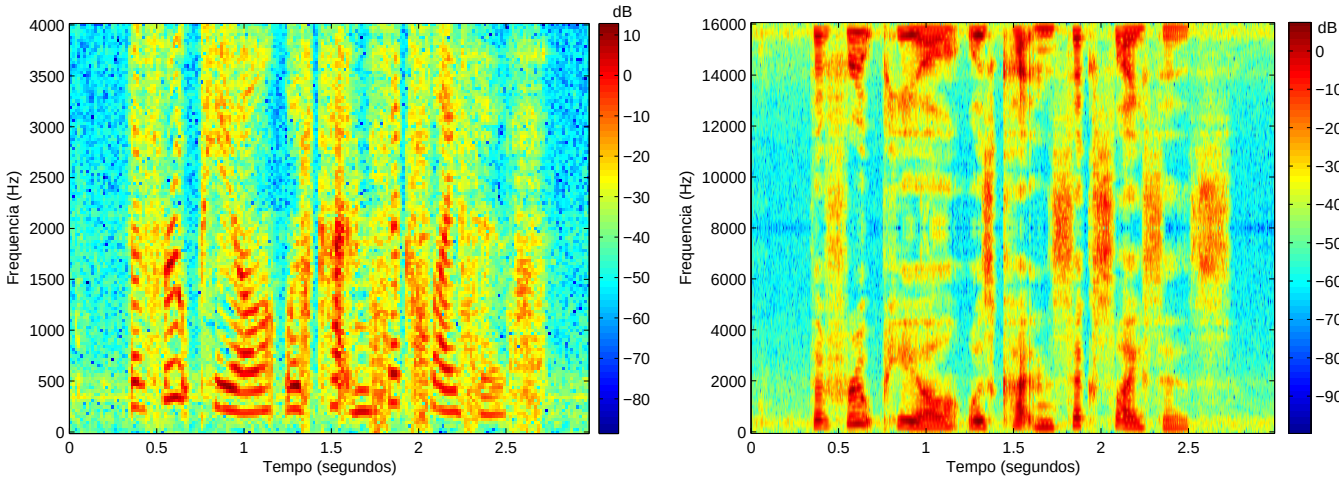
Processamento multitaxas consiste em tratar um sinal em uma ou mais freqüências de amostragem, diferentes daquela em que ele foi originalmente amostrado. Para isso é preciso que sejam realizadas operações de decimação (redução da taxa de amostragem) e/ou de interpolação (aumento da taxa de amostragem). Ambas alteram, em geral, as componentes freqüenciais do sinal, uma vez que amostras no tempo são perdidas ou inseridas no sinal original.

A fim de evitar que componentes freqüenciais sejam modificadas é preciso associar a essas operações filtros, chamados de decimadores e interpoladores. O filtro decimador é aplicado ao sinal a ser decimado $x(n)$ de acordo com a Figura 2.8(a), limitando a sua banda de freqüência e evitando, portanto, a ocorrência de *aliasing* (sobreposição de componentes freqüenciais) no sinal decimado $x_D(n)$. Já o filtro interpolador atua no sinal interpolado $x_L(n)$ de acordo com a Figura 2.8(b), eliminando as imagens geradas no espectro do sinal pela operação de interpolação.



(a) decimação por um fator M (b) interpolação por um fato L

Figura 2.8: Operações que alteram taxa de amostragem



(a) (b)

Figura 2.9: (a) Sinal de voz decimado sem filtro decimador; (b) sinal de voz interpolado sem filtro interpolador

Na Figura 2.9(a) pode-se visualizar o sinal de banda larga da Figura 1.2(a) decimado com $M = 2$ (a freqüência de amostragem, que antes era de 16 kHz, passa a ser 8 kHz). A banda do sinal decimado ocupa metade da faixa de freqüência original e, como ilustrado na Figura 2.9(a), ocorre *aliasing*. Já na Figura 2.9(b) são ilustrados os efeitos da operação de interpolação (imagens geradas em alta freqüência), também por um fator $L = 2$, do mesmo sinal da Figura 1.2(a).

Na aplicação de extensão em freqüência é interessante processar bandas de freqüência do sinal de maneiras distintas, é preciso, portanto, separá-las através de bancos de filtros que possibilitem a reconstrução perfeita do sinal a partir das bandas separadas.

Um banco de filtros é um conjunto de filtros passa faixa $H_k(z)$ (chamados filtros de análise) com entradas em comum e um conjunto de filtros passa faixa $F_k(z)$ (denominados filtros de síntese) cujas saídas são somadas para obtenção do sinal reconstruído, conceito ilustrado na Figura 2.10.

O banco de filtros é utilizado para decompor um sinal $x(n)$ em M subbandas que contêm porções diferentes do espectro do sinal $x(n)$. Existem algumas técnicas para projetar os filtros de análise e síntese garantindo reconstrução perfeita, ou seja, $y(n) = x(n - \Delta)$ caso $v'_k = v_k$, sendo Δ o atraso gerado pelos filtros. Uma delas é projetar

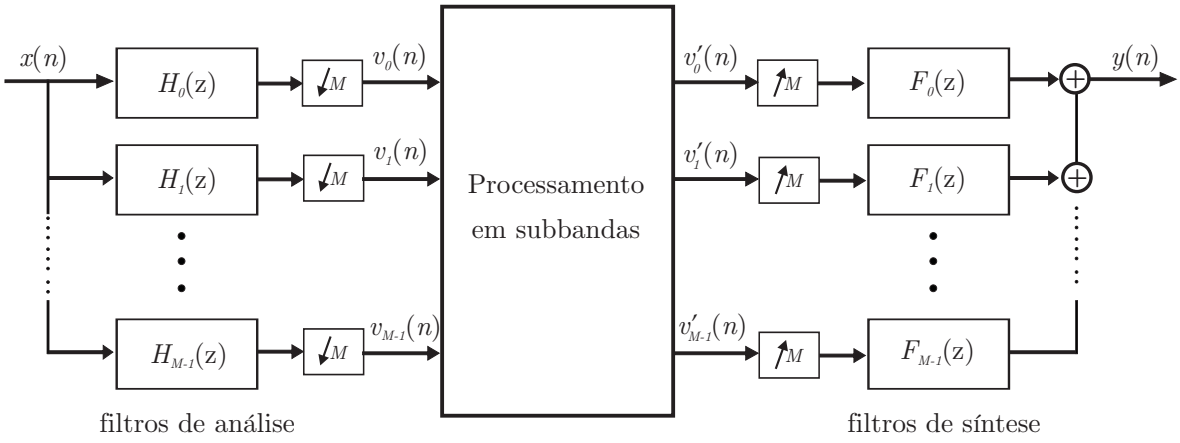


Figura 2.10: Banco de filtros de M canais

seus coeficientes, $h_k(n)$ e $f_k(n)$ respectivamente, como versões moduladas por cosseno dos coeficientes $p_0(n)$ de um filtro protótipo $P_0(z)$, ou seja:

$$h_k(n) = 2p_0(n)cos((k+\frac{1}{2})\frac{\pi}{M}(n-\frac{N}{2})+\theta_k) \tag{2.12}$$

$$f_k(n) = 2p_0(n)cos((k+\frac{1}{2})\frac{\pi}{M}(n-\frac{N}{2})-\theta_k) \tag{2.13}$$

$$\theta_k = (-1)^k\frac{\pi}{4} \tag{2.14}$$

onde N é a ordem de $P_0(z)$. Esse projeto garante que o sinal decomposto em subbandas será real, quando os coeficientes do filtro protótipo $p_0(n)$ são reais, assim como o sinal de entrada.

Considerando o filtro protótipo passa baixas cuja freqüência de corte é $\pi/2M$, com função de transferência:

$$P_o(z) = \sum_{n=0}^{N} p_o(n)z^{-n} \tag{2.15}$$

para garantir que o banco de filtros não gere distorção em amplitude significativa [9] é preciso que o filtro protótipo seja tal que a função objetivo:

$$\phi_1 = \int_0^{\pi/M} (|P_o(e^{j\omega})|^2 + |P_o(e^{j(\omega-\frac{\pi}{M})})|^2 - 1)^2 d\omega \tag{2.16}$$

seja minimizada. O que pode ser interpretado dessa função é que o módulo da distorção será suficientemente pequeno caso o termo $|P_o(e^{j\omega})|^2 + |P_o(e^{j(\omega-\frac{\pi}{M})})|^2$ seja muito próximo da unidade para toda freqüência ω.

Para garantir que não ocorra sobreposição do espectro de canais não adjacentes, o filtro protótipo deve também satisfazer:

$$|P_o(e^{j\omega})| = 0,\ \omega > \frac{\pi}{L} \tag{2.17}$$

2.3 Redes Neurais

Redes neurais artificiais tentam se comportar de maneira mais próxima possível do funcionamento do cérebro humano, através de características como capacidade de aprendizagem, generalização, abstração e robustez, e da presença de elementos organizados de tal maneira a emular a estrutura da anatomia cerebral.

A capacidade de aprendizado de uma rede neural é a característica mais vital para sua aplicação prática. Uma rede neural artificial é capaz de mudar seu comportamento a partir de variações no meio que está inserida. Ou seja, ela é capaz de se adaptar de acordo com a mudança da informação que a rede recebe.

Redes neurais artificiais também devem ser capazes de generalizar situações que lhe são apresentadas. Uma vez treinada, a resposta de uma rede até um certo ponto deve ser insensível a pequenas variações na entrada. Não é esperado, porém, que uma rede neural seja capaz de gerar novas regras sem que essas tenham lhe sido ensinadas.

A abstração de uma rede neural se traduz na sua habilidade de aprender com informações, obtendo delas o que é essencial para realização de sua tarefa (ou seja, ruídos na entrada não serão considerados) e, realizando-se um treinamento adequado, um modelo localmente ótimo será obtido.

2.3.1 Estruturas básicas de uma rede neural

Redes neurais são inspiradas em modelos biológicos, ou seja, são construídas e treinadas de acordo com pressupostos de como o cérebro humano funciona e como é estruturado. Entretanto, fazer uma analogia direta do comportamento de uma rede neural artificial com a biológica pode criar expectativas irreais sobre a capacidade de uma rede neural artificial, o que não acrescenta no desenvolvimento de pesquisas nessa área.

Tendo isso esclarecido, é vantajoso comparar, a um certo nível, elementos biológicos com os elementos que caracterizam a estrutura de uma rede neural artificial, a fim de explicar seu funcionamento e sua estrutura.

O sistema nervoso humano é constituído de células chamadas neurônios, capazes de receber, transmitir e processar um sinal eletro-químico. Os neurônios podem receber informações de outros neurônios ou de um determinado sensor e a saída pode ir, portanto, para outro neurônio ou para um atuador. Um neurônio pode receber de 10^3 até 10^4 entradas e tem somente uma saída, contínua, e ocupando uma faixa de $-50\ mV$ a $40\ mV$. A saída, mesmo sendo contínua, determina estados discretos ao neurônio, uma vez que ela é comparada a um nível para determinar se o neurônio está ativo, caso em que a saída

seja maior que esse nível, ou inativo, caso contrário.

Os contatos de entrada de um neurônio são realizados através de seus dentritos e é no axônio do neurônio que as informações vindas dos dentritos são processadas. O ato da transferência de sinal no local de conexão é chamado de sinapse nervosa. Esse contato ocorre com um ponderador chamado peso sináptico. A memória está localizada justamente nessas conexões, ou seja, nosso cérebro aprende ajustando sinapses existentes e criando outras.

Com essas considerações pode-se descrever a estrutura básica da rede neural artificial. Uma rede neural artificial é constituída de neurônios artificiais que recebem um conjunto de sinais de entrada $\{x_1 x_2 ... x_n\}$, que são as saídas de outros neurônios ou as entradas da rede. Cada entrada é multiplicada por um peso w_{ij}, correspondente ao neurônio de origem j e o neurônio de destino i, e é depois somada às outras entradas do neurônio i, resultando no sinal de ativação u_i:

$$u_i = \sum_{j=1}^{n} w_{ij} x_j + b_i = \mathbf{w}_i^T \mathbf{x} + b_i \tag{2.18}$$

no qual n é o número de entradas do neurônio i e b_i representa a polarização (*bias*), fenômeno que pode ser visto em neurônios biológicos, cujas saídas não são zero quando as entradas são zeradas.

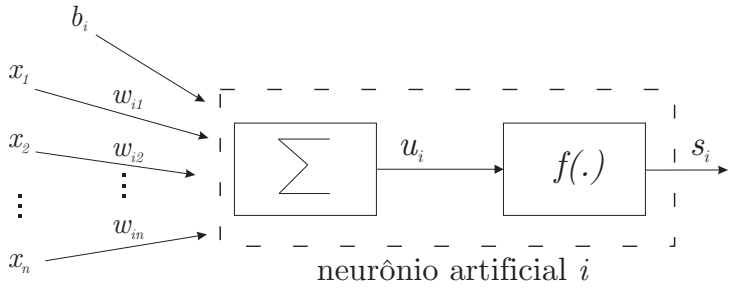
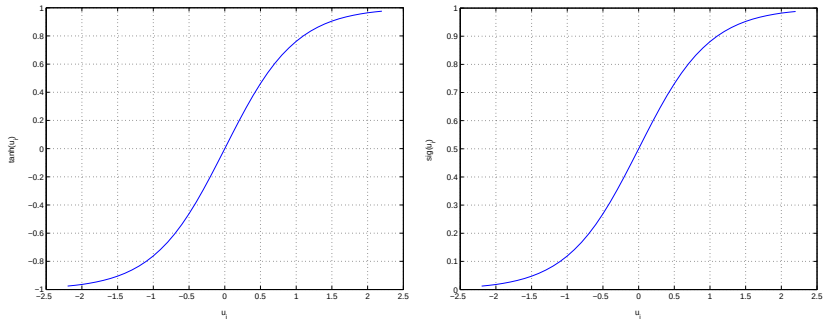


Figura 2.11: Estrutura de um neurônio artificial

A saída s_i do neurônio artificial é o sinal u_i após ser modificado pelo que se denomina função de ativação $f(\cdot)$ do neurônio. Pode-se empregar os seguintes tipos de funções de ativação:

- linear: excitação interna será a saída;
- não linear: dentro desse caso é possível citar, dentre outras funções, a tangente hiperbólica $tanh(u_i)$ e a sigmóide $sig(u_i)$, que seguem as seguintes definições:

$$tanh(u_i) = \frac{1 - e^{-2u_i}}{1 + e^{-2u_i}} = \frac{2}{1 + e^{-2u_i}} - 1 = 2\ sig(u_i) - 1 \tag{2.19}$$



(a) (b)

Figura 2.12: Funções de ativação não lineares: (a) tangente hiperbólica; (b) sigmóide

- lógica: descrita da seguinte maneira:

$$f(u_i) = \begin{cases} 1, & u_i \geq 0 \\ 0, & u_i < 0 \end{cases} \tag{2.20}$$

Quando é esperada a implementação de operações mais complexas, os neurônios devem ser arranjados em redes. A configuração de uma rede mais simples contém somente uma camada de neurônios. Os neurônios de uma rede neural artificial são, portanto, arrumados por camadas e a informação flui somente em um sentido no caso das redes *feedforward*, que são as redes utilizadas neste projeto. Essa característica estrutural garante a estabilidade da rede neural, uma vez que não existem ramos de realimentação.

Uma rede de uma camada, porém, não é capaz de resolver problemas simples como o "ou-exclusivo" [10], já uma rede de duas camadas é capaz de resolver potencialmente qualquer problema de classificação. A quantidade de neurônios na última camada necessariamente será igual à quantidade de saídas da rede. Portanto, o dimensionamento de uma rede envolve a escolha dos seguintes parâmetros:

- número de neurônios N na camada intermediária, dita camada escondida
- tipo de neurônio utilizado em cada camada

Redes neurais são capazes de resolver problemas de classificação e de mapeamento. Uma rede classificatória deve possuir necessariamente duas camadas para conseguir separar classes que não sejam linearmente separáveis dentro do plano definido pelas entradas da rede. Uma classe é dita linearmente separável quando é possível separá-la de outras classes no plano-N através de somente um elemento de $N-1$ dimensões. Ou seja, caso a rede possua duas entradas, uma classe linearmente separável deve poder ser separada das outras classes por uma reta. A primeira camada de uma rede classificatória atua nas classes de tal maneira a encontrar uma nova representação em que todas as classes estejam linearmente separáveis. A segunda camada atua, então, separando essas classes nesse novo plano formado pelas saídas dos neurônios da primeira camada.

2.3.2 Aplicações de redes neurais

Uma vez vistas as características das redes neurais artificiais, é possível prever em quais aplicações o seu uso seria de melhor proveito. É certo discernir duas situações em que redes neurais artificiais se apresentam como a melhor alternativa para resolução de um problema. Uma delas é quando não existe um modelo fenomenológico aceitável para um sistema, ou seja, quando não é possível encontrar modelos matemáticos que descrevem o sistema de maneira a atingir um certo nível de excelência dada uma métrica. O outro caso é quando não existe nenhum modelo conhecido capaz de representar o fenômeno.

A extensão em freqüência é um clássico problema onde redes neurais podem ser aplicadas, já que não existe um modelo matemático que propriamente descreva o mapeamento entre as freqüências perdidas e aquelas que continuam presentes no sinal filtrado. O projeto desenvolvido demonstrou a capacidade da rede neural de mapear relações não lineares presentes nas características da voz humana, atuando como uma rede classificatória e mapeadora.

2.3.3 Treinamento

Treinar uma rede neural artificial é ajustar os valores dos pesos e *bias* de cada sinapse. Existem duas maneiras de se treinar uma rede neural: a supervisionada e a não supervisionada. Em um treinamento não supervisionado não possuímos a informação da saída que a rede deve encontrar dada uma entrada. Já em um treinamento supervisionado os pesos são ajustados de acordo com algum método que leve em consideração o erro gerado pela saída da rede, já que temos a informação de qual seria a saída esperada. A informação pode ser qual a classe a que determinado dado de entrada pertence, no caso de redes classificatórias, ou qual saída deve ocorrer quando aplicamos certa entrada, no caso de redes mapeadoras.

Um método de treinamento amplamente utilizado é o *backpropagation* que consiste em ajustar os valores dos pesos de acordo com erros que são propagados da saída da rede até a sua entrada. Para isso é necessário que os neurônios utilizem funções de ativação diferenciáveis em relação à entrada u, o que explica o uso das funções lineares e não-lineares descritas na Seção 2.3.1.

Existem diversas variações do *backpropagation*, uma delas é o treinamento gradiente descendente realizado em batelada. Nesse tipo de treinamento os pesos e valores de *bias* são atualizados na direção negativa do gradiente da função de performance somente depois que um conjunto de pares de entrada e saída for utilizado para o cálculo do gradiente. Os gradientes calculados para cada par de entrada e saída do conjunto são então somados para determinar qual o gradiente que será utilizado na expressão de atualização dos coeficientes w_{ij}:

$$\mathbf{W} = \mathbf{W} + \Delta\mathbf{W} \tag{2.21}$$

sendo $\mathbf{W} = [\mathbf{w}_1\mathbf{w}_2 \ ... \ \mathbf{w}_N]^T$ e $\mathbf{w}_i$ o vetor com os pesos das sinapses do neurônio i,

$$\Delta\mathbf{W} = -\alpha\nabla_{\mathbf{W}}F_o \tag{2.22}$$

$$\nabla_{w_{ij}}F_o = \frac{\partial F_o}{\partial w_{ij}} \tag{2.23}$$

F_o é a função que será minimizada no ajuste dos pesos, que pode ser por exemplo o erro médio quadrático:

$$F_o = E(\varepsilon^{2^k}) = \frac{1}{P}\sum_{k=1}^{P}\varepsilon_k^2 \tag{2.24}$$

sendo P o número de pares de entrada e saída que é utilizado para treinar a rede e ε_k^2 a soma dos erros quadráticos de cada saída do k-ésimo par.

Existem alguns parâmetros que devem ser ajustados quando esse treinamento é utilizado, como o número de épocas, ou seja, o número de conjuntos em que os pares de entrada e saída serão organizados para realizar o treinamento; o objetivo, valor que diz quando o treinamento deve ser interrompido de acordo com a função de performance escolhida, e a taxa de aprendizagem α, que é o passo que será dado em direção oposta ao gradiente, de acordo com a Equação (2.22).

Quando o fenômeno não está suficientemente representado pelos pares de entrada e saída utilizados no treinamento, este pode vir a gerar ajustes excessivos nos pesos e valores de *bias*. Isso faz com que a rede perca sua capacidade de generalização, o que pode ser evitado se juntamente com o treino for realizada a validação da rede treinada. Ou seja, a cada N_e épocas o treino cessará e a rede será testada com outros pares de entrada e saída, que não serão utilizados no treinamento. Dessa maneira o treinamento irá cessar caso o erro de validação continue crescendo após K testes de validação implementados. Com isso, obtém-se outro critério para interromper o treinamento da rede.

2.3.4 Qualidade de um classificador

Para avaliar a qualidade de um classificador existem diversos parâmetros. Considerando uma rede que discrimine duas classes, os vozeados e os não vozeados, pode-se enumerar quatro possíveis situações:

- V_V - verdadeiros vozeados, ou seja, quando a rede classifica como vozeada uma entrada vozeada;
- F_V - falsos vozeados, ocorre quando a rede classifica como vozeada uma entrada não vozeada;
- V_{NV} - verdadeiros não vozeados, quando a rede classifica como não vozeada uma entrada não vozeada;

- F_{NV} - falsos não vozeados, ocorre quando a rede classifica como não vozeada uma entrada que é vozeada.

Tem-se portanto que os erros ocorrem nas situações F_V e F_{NV}. A taxa de acerto dessa rede é assim definida:

$$T_a = \frac{V_V + V_{NV}}{V_V + V_{NV} + F_V + F_{NV}} \tag{2.25}$$

2.4 Filtragem Adaptativa

A teoria de filtragem adaptativa tem como objetivo encontrar um filtro capaz de gerar uma saída $y(n)$ que se aproxime do sinal desejado $d(n)$ dado o sinal de entrada $x(n)$ [11], conceito ilustrado na Figura 2.13. Caso o sinal desejado seja um ruído branco, ou seja, sem correlação entre suas amostras, o filtro encontrado será sub-ótimo se comparado com o filtro de Wiener, cujos coeficientes são obtidos através do conhecimento da estatística dos sinais envolvidos. Porém, os algoritmos de filtragem adaptativa são capazes de rastrear melhor as variações contidas nos sinais $x(n)$ e $d(n)$.

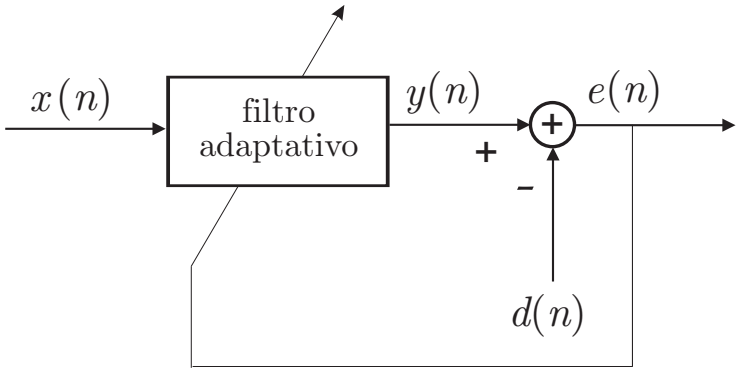


Figura 2.13: Estrutura do conceito de filtragem adaptativa

Existem diversos algoritmos de filtragem adaptativa, dentre os quais podemos citar o método *Steepest-Descent*, que utiliza o conhecimento estatístico dos sinais envolvidos para encontrar o filtro ótimo, sendo a sua função de custo o valor esperado do erro médio quadrático.

Outro algoritmo amplamente utilizado é o LMS (*Least Mean Square*), que usa estimativas instantâneas das estatísticas dos sinais para minimizar a mesma função custo do método anterior. É, portanto, uma alternativa com baixa complexidade computacional e mais sensível a variações dos sinais $x(n)$ e $d(n)$ que o método *Steepest-Descent*, porém apresenta uma certa variação em torno do filtro ótimo, que depende do valor do passo de adaptação. A estabilidade de ambos algoritmos deve ser analisada para escolher o passo de adaptação.

Uma alternativa aos métodos que atualizam o filtro encontrado a cada amostra nova dos sinais $x(n)$ e $d(n)$ é o método dos mínimos quadrados, que encontra o filtro ótimo para um certo intervalo de amostras já conhecido desses sinais.

2.4.1 Método dos mínimos quadrados

O método dos mínimos quadrados se diferencia de outros métodos pela função custo que será minimizada:

$$\varepsilon = \sum_{n=M_1}^{M_2} e^2(n) \tag{2.26}$$

na qual:

$$\begin{aligned} e(n) &= d(n) - y(n) \\ &= d(n) - \sum_{k=0}^{M-1} w_k x(n-k) \end{aligned} \tag{2.27}$$

Para que esse método seja aplicável é preciso conhecer as amostras $\{x(1), x(2), ...x(N)\}$ e $\{d(1), d(2), ...d(N)\}$ dos sinais de entrada e desejado, respectivamente. Uma boa escolha para o intervalo de amostras do erro a ser considerado na função custo é $M_1 = M$ (M é a ordem do filtro adaptativo) e $M_2 = N$, de acordo com o método das covariâncias [11].

Para encontrar o filtro que minimize a função de custo apresentada na Equação (2.26) é preciso que:

$$\frac{\partial \varepsilon}{\partial w_k} = 0 \tag{2.28}$$

Dessa condição encontra-se o que é chamado de equações normais determinísticas:

$$\sum_{n=M}^{N} x(n-k)\mathbf{e}_o(n) = 0 \tag{2.29}$$

ou

$$X^T \mathbf{e}_o = 0 \tag{2.30}$$

onde

$$\mathbf{e}_o = [e_o(M)...e_o(N)] \tag{2.31}$$

$$X^T = \begin{bmatrix} x(M) & x(M+1) & \dots & x(N) \\ x(M-1) & \ddots & & x(N-1) \\ \vdots & & & \vdots \\ x(1) & x(2) & \dots & x(N-M+1) \end{bmatrix} \tag{2.32}$$

Nas equações normais determinísticas é visto que o vetor com os erros $\mathbf{e}_o$ obtido com o filtro ótimo é ortogonal a cada coluna da matriz com os dados de entrada X.

Os coeficientes ótimos são então descritos pela equação:

$$w_o = [X^T X]^{-1}[X^T d] \tag{2.33}$$

na qual $d^T = [d(M)d(M+1)...d(N)]$.

Capítulo 3

Métodos para extensão em freqüência

Os métodos de extensão em freqüência aqui propostos realizam o processamento nas sub-bandas do sinal de telefone limitado em 4 kHz e amostrado a 16 kHz ($s_{tel16k}(n)$). Cada sub-banda, portanto, é estendida separadamente de acordo com as técnicas propostas nos dois métodos. Antes de apresentá-los será descrito o que ambos realizam em comum, como a separação do sinal $s_{tel16k}(n)$ em 16 canais.

O sinal $s_{tel}(n)$, que corresponde ao sinal de banda estreita amostrado a 8 kHz, é inicialmente interpolado por 2, gerando o $s_{tel16k}(n)$, que é ainda limitado em 4 kHz, mas com taxa de amostragem igual a 16 kHz. Este sinal é dividido em 16 canais, através de um banco de filtros modulado por cosseno [9], cada canal contendo, portanto, uma banda de frequência de aproximadamente 500 Hz de largura deste sinal. As 16 bandas que são separadas do sinal $s_{tel16k}(n)$ podem ser visualizadas na Figura 3.2. As primeiras sete bandas $\{s_1(n)\ s_2(n)\ ...\ s_7(n)\}$, que correspondem à faixa de frequência de 0 a 3500 Hz do sinal, não são modificadas e irão fazer parte do sinal reconstruído. As outras nove bandas, que correspondem às frequências de 3500 a 8000 Hz, serão estimadas $\{\tilde{s}_8(n)\ \tilde{s}_9(n)\ ...\ \tilde{s}_{16}(n)\}$ a partir dos componentes de baixas frequências do sinal $s_{tel16k}(n)$.

Foi utilizado como filtro protótipo para o banco de filtros, o filtro de 256 coeficientes proposto em [12] [13]. As respostas em freqüência dos 16 filtros de análise $\{H_0(z)H_1(z)...H_{15}(z)\}$, que foram encontrados utilizando este protótipo, podem ser visualizadas na Figura 3.1.

3.1 Algoritmo 1

O primeiro algoritmo proposto realiza a estimação das componentes de mais altas freqüências através de um mapeamento linear, implementado por filtros lineares com resposta ao impulso finita (FIR), que são escolhidos dentre dois conjuntos de filtros (previamente calculados) de acordo com a classificação de cada trecho do sinal. A estrutura do algoritmo pode ser vista na Figura 3.3.

O método proposto é computacionalmente mais simples que o método que utiliza

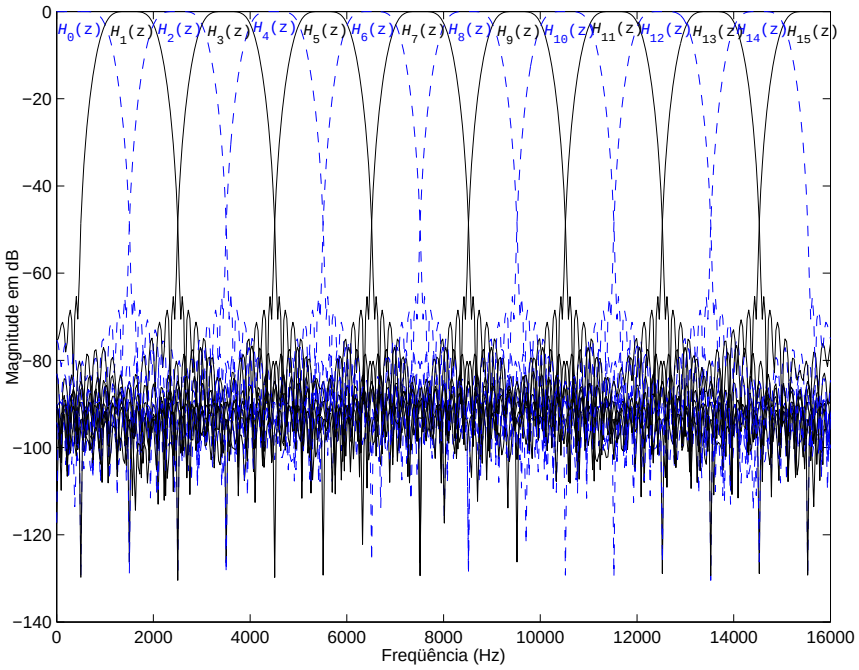


Figura 3.1: Filtros de análise do banco de filtros modulado por cosseno de 16 canais utilizado em ambos algoritmos

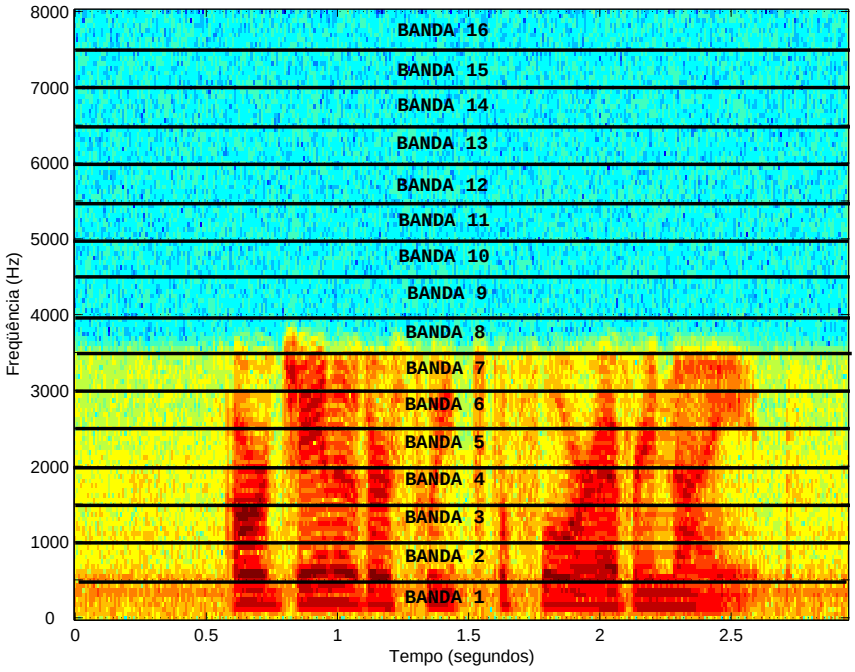


Figura 3.2: Sinal $s_{tel16k}(n)$ dividido em 16 bandas de freqüência

um *codebook*, pois ao invés de tratar individualmente cada sinal de voz, este é discriminado em duas grandes classes: fonemas vozeados e não vozeados, já discutidas no Capítulo 1.

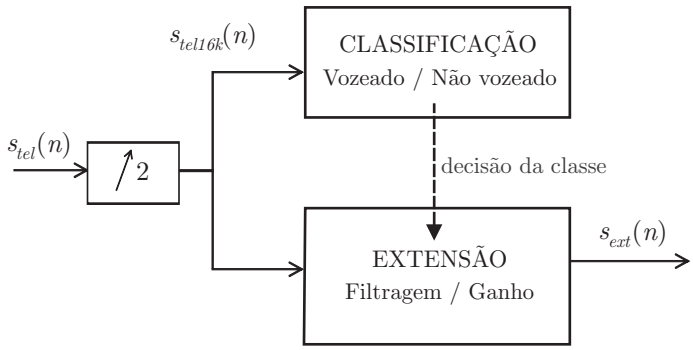


Figura 3.3: Estrutura do Algoritmo 1

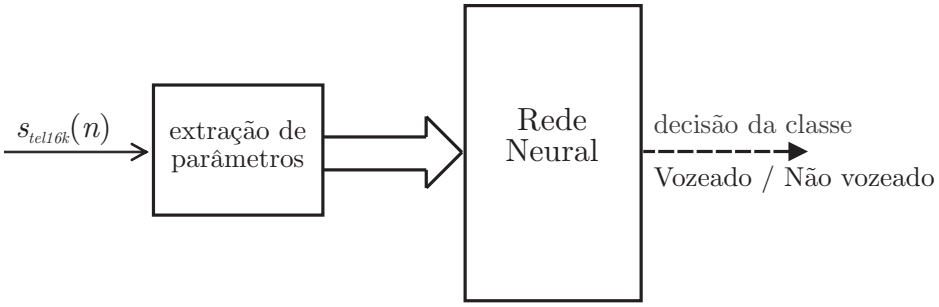


Figura 3.4: Estrutura da classificação realizada pelo Algoritmo 1

A classificação é implementada através de uma rede neural cujas entradas são 5 parâmetros escalares extraídos do sinal $s_{tel16}(k)$, de acordo com a Figura 3.4. Os parâmetros propostos em [14] são:

- autocorrelação para retardo igual a 1
- autocorrelação para retardo igual ao intervalo do *pitch*
- *zero crossing rate*, taxa de troca de sinal do $s_{nb}(n)$ (trecho de 20 ms (320 amostras) do sinal $s_{tel16}(n)$).
- *kurtosis* local, definido como:

$$x_{lk} = \log \frac{1}{N} \sum_{n=0}^{N-1} s_{nb}^4(n) - 2 \log \frac{1}{N} E(m) \tag{3.1}$$

sendo $E(m) = \sum_{n=0}^{N-1} s_{nb}^2(n)$

- centróide espectral, calculado pela expressão:

$$x_{sc} = \frac{\sum_{i=0}^{N/2} i \cdot |S_{nb}(e^{j\omega_i})|}{(\frac{N}{2}+1)\sum_{i=0}^{N/2} |S_{nb}(e^{j\omega_i})|} \tag{3.2}$$

sendo $S_{nb}(e^{j\omega_i})$ o i-ésimo termo da transformada discreta de Fourier (DFT) do sinal $s_{nb}(n)$.

Como fonte de dados para o sistema, foram utilizadas seis frases em inglês com cerca de 3 segundos de duração cada, sendo três vozes masculinas e três femininas. O treinamento da rede foi realizado com alguns trechos de 20 ms destes sinais de voz, recortados e divididos entre vozeados e não vozeados. Para o treinamento da rede, foram separados 940 pares, e para validação foram utilizados 415 pares, sendo que desse total 664 eram vozeados e 691 não vozeados.

A Figura 3.5 apresenta o erro médio quadrático (MSE) obtido no treinamento e na validação da rede do tipo *feedforward* utilizada com a seguinte estrutura: 4 neurônios na camada escondida e 1 neurônio na camada de saída, função de ativação tangente hiperbólica para cada camada e algoritmo de treinamento *backpropagation*, de gradiente descendente [10].

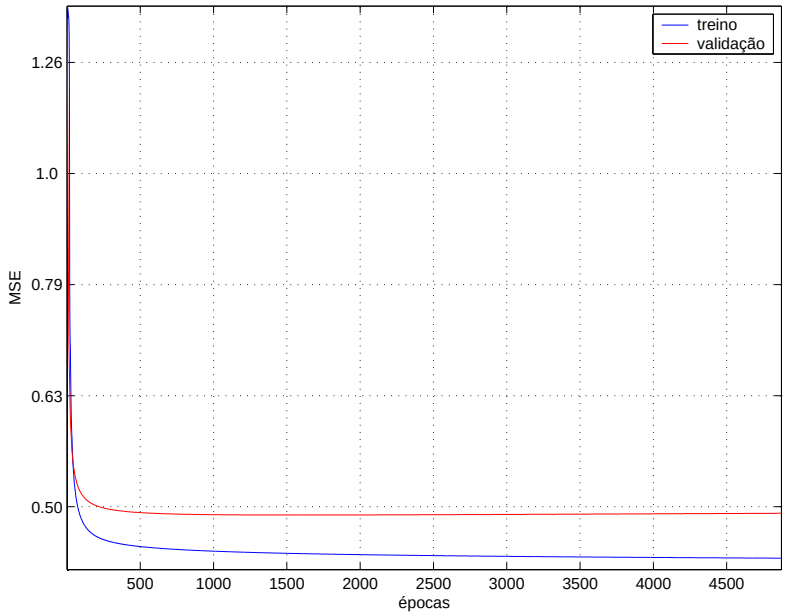


Figura 3.5: Erro médio quadrático computado no treino e na validação da rede neural do Algoritmo 1

Procurou-se treinar a rede de forma a encontrar pesos que resultassem na máxima discriminação entre as classes para um número mínimo de neurônios, através da análise das componentes principais [15].

A rede, formada inicialmente por um único neurônio na camada escondida e um neurônio na camada de saída, é treinada pelo método de *backpropagation*. A validação dos dados foi realizada como critério de parada. Os pesos encontrados para o neurônio da

N_n	MSE final	T_a
1	0,4617	0,94118
2	0,4521	0,96471
3	0,4516	0,96471
4	0,4505	0,96471

Tabela 3.1: Valores de MSE final e taxa de acerto T_a obtidos no treinamento de redes com diferentes números de neurônios N_n na camada escondida

camada escondida neste treinamento representam a primeira componente de discriminação extraída. Uma nova rede é gerada a partir da adição de um neurônio nesta camada. Todos os pesos desta rede são então treinados e o vetor de pesos que conecta o espaço de entrada ao novo neurônio irá gerar a segunda componente. Desta maneira novas componentes são extraídas até que a taxa de acerto e o erro médio quadrático (MSE) final não se alterem significativamente, o que ocorreu a partir da criação de redes com mais de quatro neurônios. Os valores do MSE final e da taxa de acerto T_a obtidos para cada rede gerada estão dispostos na Tabela 3.1. Foi verificado que o treinamento dos neurônios inicializados com componentes previamente extraídas não gerou ajustes significativos. Para testar as redes foram utilizados 340 pares do conjunto utilizados para treino e validação.

Após a identificação da classe a que o trecho de voz processado pertence, o sinal $s_{tel16k}(n)$ é decomposto em 16 bandas com a utilização do banco de filtros modulado por cosseno.

O sinal da sétima banda $s_7(n)$, considerada a banda de maior correlação com as bandas perdidas [1], será utilizado como entrada para os filtros ótimos $H_{o_k}(z)$ encontrados para cada uma das 8 bandas de maiores freqüências (k variando de 8 até 16), de acordo com a Figura 3.6. A saída $\tilde{s}_k(n)$ do k-ésimo filtro gera a k-ésima banda estimada que será utilizada, juntamente com as outras bandas estimadas e com as bandas de 1 a 7 do sinal s_{tel16k}, no banco de filtros de síntese, para gerar o sinal reconstruído $s_{ext}(n)$.

Os filtros ótimos $H_{o_k}(z)$ foram obtidos pelo método dos mínimos quadrados [6] para cada uma das duas classes. O sinal desejado para o filtro da k-ésima banda $H_{o_k}(z)$ é o sinal da k-ésima banda $s_k(n)$ da voz de banda larga, e a entrada para os filtros de todas as sub-bandas é o sinal da sétima banda $s_7(n)$, como ilustrado na Figura 3.6.

Foram realizados testes com filtros de diferentes ordens. Concluiu-se que, para os fonemas não vozeados, filtros de comprimento 8 geram resultados satisfatórios, enquanto que para os fonemas vozeados, um único coeficiente foi suficiente para relacionar as componentes das bandas de mais altas freqüências com as da sétima banda. Então, de acordo com a classificação do trecho processado, uma filtragem ou um ganho é aplicado sobre a sétima banda do sinal de banda estreita para a geração das outras 8 bandas. Na Figura 3.7 encontram-se as respostas em freqüência dos filtros calculados para as nove bandas superiores do sinal $s_{tel16k}(n)$, e na Figura 3.8 são mostrados os valores dos ganhos aplicados

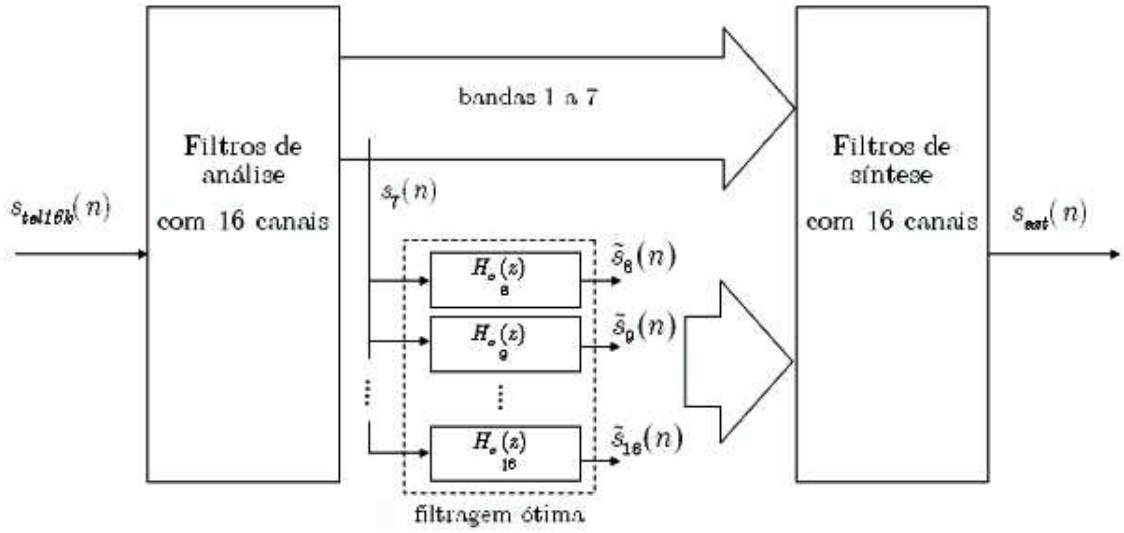


Figura 3.6: Estrutura da extensão implementada pelo Algoritmo 1

ao sinal da sétima banda para gerar o sinal estendido para fonemas vozeados.

3.2 Algoritmo 2

O segundo algoritmo proposto realiza a estimação do envelope espectral $H_{AR_k}(z)$ da k-ésima banda do sinal $s_{tel16k}(n)$, com k variando de 8 a 16, através de um modelo AR de primeira ordem para cada banda.

Uma rede neural é utilizada para gerar os coeficientes $\tilde{a}_{1_k}$ do modelo AR das 9 últimas bandas de freqüência do sinal $s_{tel16k}(n)$ a partir dos coeficientes a_{1_b}, b variando de 1 a 7, obtidos para as primeiras 7 bandas de freqüência do sinal de banda estreita, isto é, de 0 a 3500 Hz, conforme ilustrado na Figura 3.9. Estes são os coeficientes do filtro do erro de predição (LPC) que compõem o filtro do trato vocal $H_{TV}(z)$ discutido na Subseção 2.1.4. Temos então:

$$H_{TV_k}(z) = H_{AR_k}(z) = \frac{c_k}{1 + \tilde{a}_{1_k} z^{-1}} \tag{3.3}$$

sendo c_k um ganho calculado de acordo com a seguinte expressão:

$$c_k = b_k \sqrt{\frac{(1 + \tilde{a}_{1_k})^2}{(1 + a_{1_7})^2}} \tag{3.4}$$

na qual $0.1 \leq b_k \leq 0.9$ é um peso ajustado experimentalmente para cada subbanda. Este ajuste foi feito com o objetivo de evitar artefatos e ruídos de alta freqüência, portanto, o valor do peso das bandas de mais alta freqüência é feito menor.

O sinal de excitação utilizado para reconstruir todas as bandas superiores foi o erro de predição encontrado para a sétima banda $e_7(n)$, como mostrado na Figura 3.10.

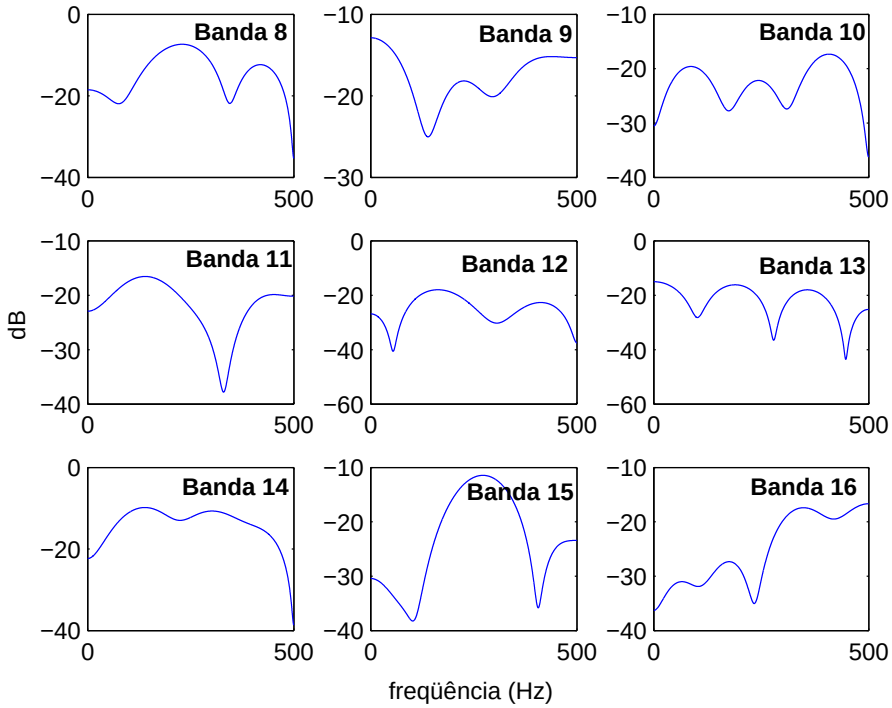


Figura 3.7: Resposta em freqüência dos filtros para estender as nove bandas superiores de sinais não vozeados.

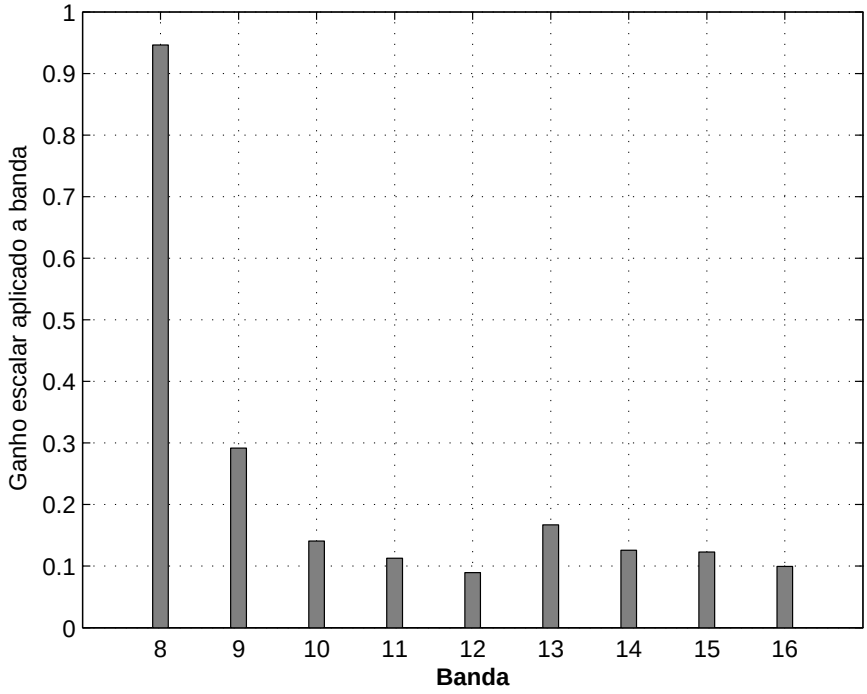


Figura 3.8: Ganhos aplicados para estender as nove bandas superiores de sinais vozeados.

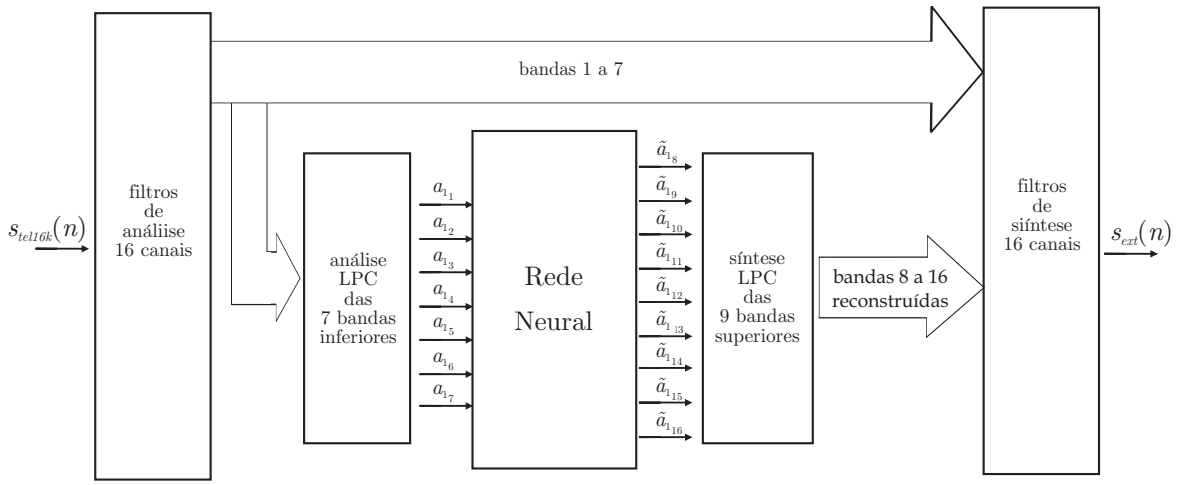


Figura 3.9: Estrutura do Algoritmo 2.

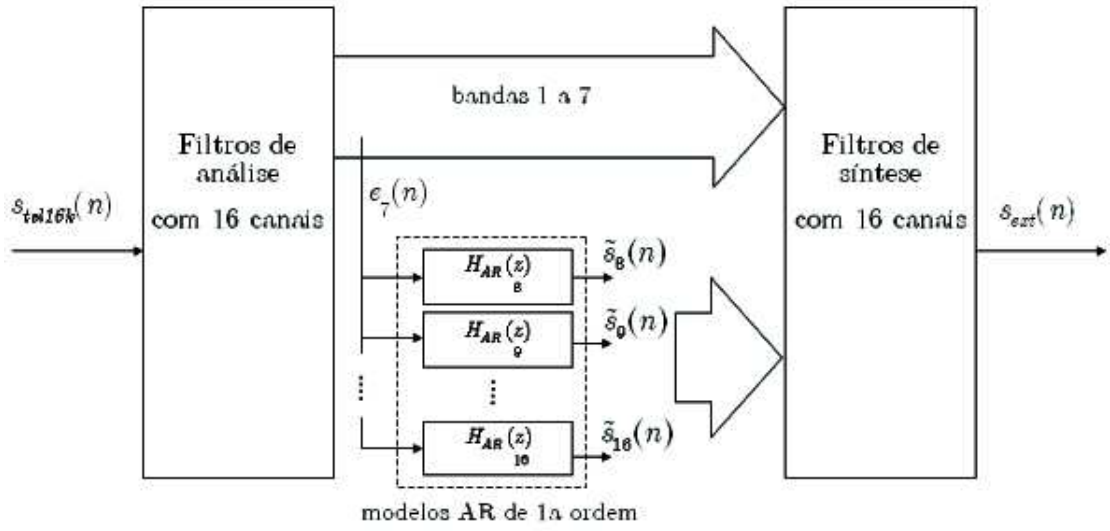


Figura 3.10: Estrutura da extensão implementada pelo Algoritmo 2.

Como fonte de dados para o treinamento e teste da rede neural, foram utilizadas as mesmas seis frases em inglês do Algoritmo 1. Estas frases foram repartidas em amostras com duração de 20 ms cada. Em seguida, de cada amostra foi extraído o primeiro coeficiente a_1 do filtro AR modelado para as 16 bandas do sinal. Para a entrada da rede foram utilizados os coeficientes das 7 primeiras bandas $\{a_{1_1}\ a_{1_2}\ ...\ a_{1_7}\}$ das frases do sinal limitado e para a saída da rede foram utilizados os coeficientes das 9 últimas bandas $\{a_{1_8}\ a_{1_9}\ ...\ a_{1_{16}}\}$ das frases originais de banda larga. A rede utilizada foi uma rede *feedforward* com neurônios do tipo tangente hiperbólica na camada escondida e linear na última camada. O algoritmo de treinamento escolhido foi o *back-propagation* de gradiente descendente. Os melhores resultados foram obtidos com 9 neurônios na camada intermediária.

A base de dados disponível para o sistema consistia em 800 pares de entrada e saída.

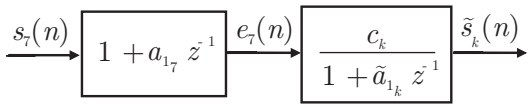


Figura 3.11: Extensão realizada pelo algoritmo 2 na k-ésima banda

Para o treinamento da rede, foram separados 600 pares, e para validação foram utilizados 200 pares. A Figura 3.12 apresenta o erro médio quadrático (MSE) obtido no treinamento e validação da rede.

Tanto os dados de entrada quanto os de saída foram normalizados, de modo que o módulo máximo de cada um fosse unitário. Em seguida, foi extraída a correlação cruzada entre as sete entradas para revelar a dependência entre as variáveis de entrada. Feito isso, foram extraídas as correlações cruzadas entre as sete entradas e as nove saídas da rede. Com os resultados, expostos nas Tabelas 3.2 e 3.3, pode-se observar que as entradas correspondentes às bandas dois e quatro não tiveram forte influência na saída. Porém, como o número de dados não era grande, os dados de entrada referentes a essas duas bandas não foram descartados. A entrada correspondente à sétima banda teve, em geral, maior correlação com as saídas.

	1	2	3	4	5	6	7
1	1	1,65E-12	0,57311	-0,001742	0,075487	-0,40611	-0,59149
2	1,65E-12	1	1,62E-12	5,20E-16	2,92E-13	8,74E-13	-5,52E-13
3	0,57311	1,62E-12	1	0,000872	0,14464	-0,41972	-0,89078
4	-0,001742	5,20E-16	0,000872	1	0,000221	0,000971	-0,000221
5	0,075487	2,92E-13	0,14464	0,000221	1	-0,03441	-0,12284
6	-0,40611	8,74E-13	-0,41972	0,000971	-0,03441	1	0,69045
7	-0,59149	-5,52E-13	-0,89078	-0,000221	-0,12284	0,69045	1

Tabela 3.2: Correlações entre as 7 entradas da rede neural

	1	2	3	4	5	6	7
1	-0,59028	-1,44E-13	-0,83429	-3,17E-01	-0,11329	0,72678	0,98725
2	-0,12352	1,28E-12	-0,49464	-0,00017926	-0,051501	0,87045	0,75874
3	-0,54415	2,21E-13	-0,46609	0,00077711	-0,044625	0,82572	0,58666
4	0,5587	1,27E-13	0,96401	0,0010166	0,14987	-0,43151	-0,91951
5	-0,54866	-3,55E-13	-0,87604	-0,00035905	-0,12112	0,75064	0,97633
6	-0,54055	1,27E-14	-0,72348	0,00043655	-0,086777	0,82081	0,94656
7	-0,72404	-7,83E-13	-0,87697	0,00025001	-0,11878	0,69403	0,98397
8	0,59703	8,17E-13	0,93201	0,0004763	0,13402	-0,61492	-0,9912
9	-0,09004	1,36E-12	0,41119	0,0022648	0,088645	0,24901	-0,15787

Tabela 3.3: Correlações entre as 7 entradas e as 9 saídas da rede neural

Como não há garantia que os coeficientes gerados pela rede implementem filtros só

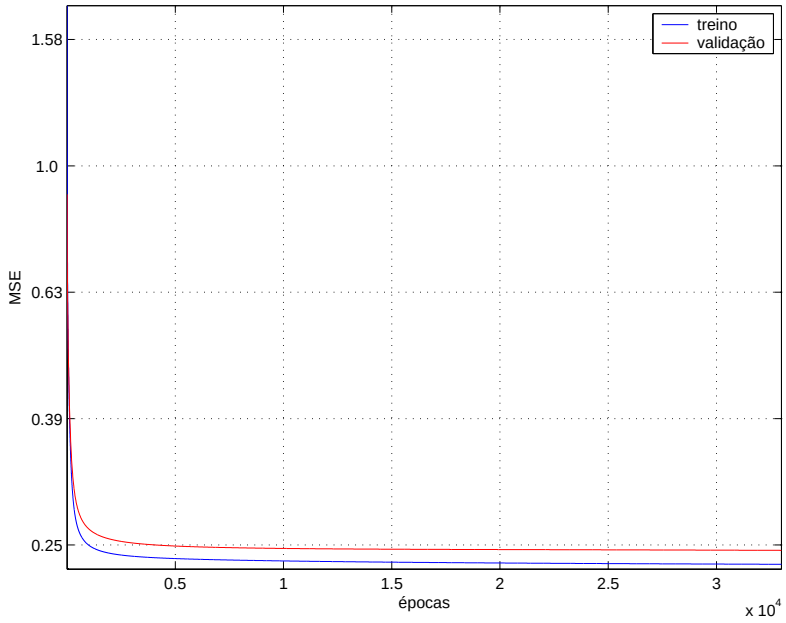


Figura 3.12: Erro médio quadrático computado no treino e na validação da rede neural do Algoritmo 2

pólos estáveis para o modelo do trato vocal de cada banda, é preciso checar a estabilidade dos coeficientes da k-ésima banda da seguinte maneira:

$$\text{se} \quad \tilde{a}_{1_k} > 1 \text{ então} \quad \tilde{a}_{1_k} = 1 - (\tilde{a}_{1_k} - 1)$$

Assim os pólos instáveis serão refletidos dentro do círculo unitário [4], garantindo a estabilidade do filtro $H_{AR_k}(z)$.

A estrutura da extensão realizada na k-ésima banda é ilustrada na Figura 3.11.

Capítulo 4

Resultados e análises

Neste capítulo serão apresentados os resultados obtidos com ambos algoritmos, assim como a análise quantitativa e subjetiva desses resultados. As frases utilizadas para tal estão expostas na Tabela 4.1.

Nome de referência	Frase
Mulher 1	*Our janitor sweeps the floor every night*
Mulher 2	*There isn't enough paint to finish the room*
Mulher 3	*The fruit peel was cut in six slices*
Homem 1	*Use a pencil to write the first draft*
Homem 2	*If your tooth hurts that much you should see a dentist*
Homem 3	*Tuck the sheet under the edge of the mat*

Tabela 4.1: Frases pronunciadas por cada locutor

4.1 Resultados

Os resultados obtidos para os dois algoritmos propostos serão expostos através dos espectrogramas dos sinais reconstruídos para cada frase.

Ilustradas nas Figuras 4.1, 4.2, 4.3, 4.4, 4.5 e 4.6 encontram-se as classificações (NV: não vozeado e V: vozeado) realizadas pela rede neural do Algoritmo 1 para as seis frases testadas, sobrepostas ao espectrograma do sinal de banda larga correspondente para que ocorra uma avaliação mais clara dos acertos e erros da rede.

Na Figura 4.1 é visto que os fonemas da frase Mulher 1 no começo da palavra *janitor* e no começo e final da palavra *sweeps* são classificados como não vozeados, assim como o /th/ da palavra *the*, o f da palavra *floor*, o r de *every* e o final da palavra *night*. Já na frase Homem 1 a rede classificou como não vozeado trechos como o final da palavra *use*, o fonema relacionado à letra *c* da palavra *pencil* e os fonemas gerados pela palavra *to* e pelas letras f, e st de *draft* e o inicio da palavra *draft*.

É visto, portanto, que tanto para vozes femininas quanto para masculinas a rede

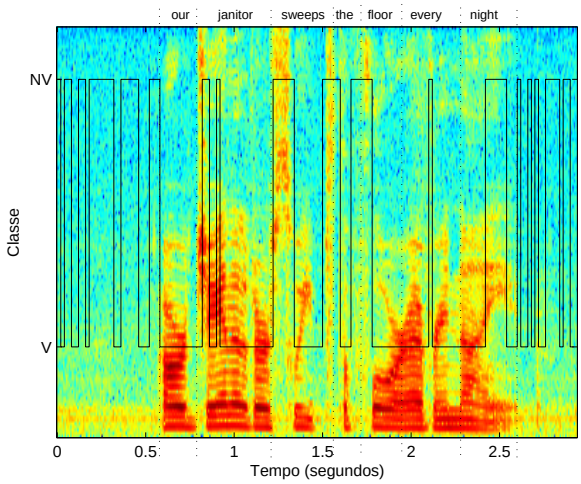


Figura 4.1: Classificação entre NV (não vozeado) e V (vozeado) para Arquivo Mulher 1

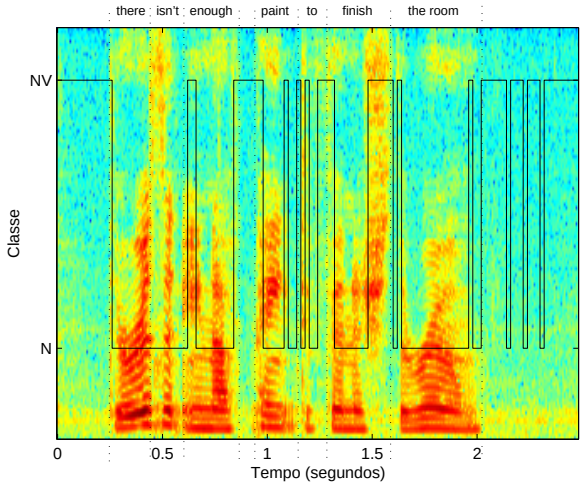


Figura 4.2: Classificação entre NV (não vozeado) e V (vozeado) para Arquivo Mulher 2

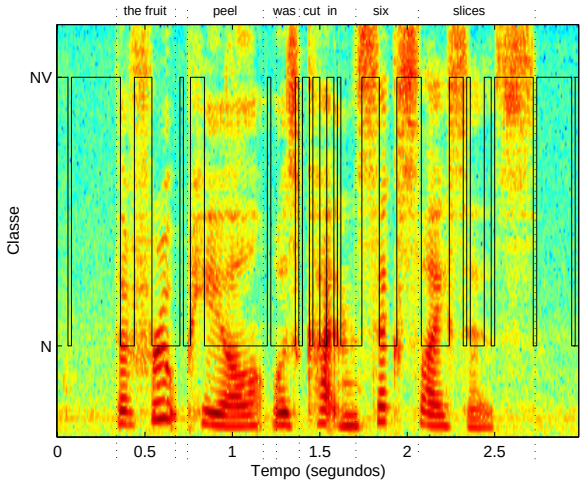


Figura 4.3: Classificação entre NV (não vozeado) e V (vozeado) para Arquivo Mulher 3

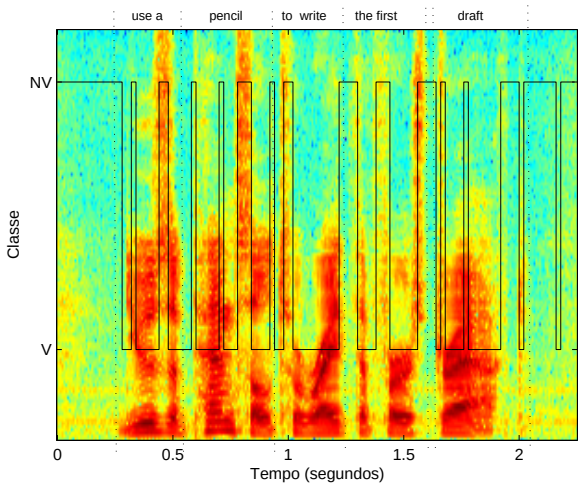


Figura 4.4: Classificação entre NV (não vozeado) e V (vozeado) para Arquivo Homem 1

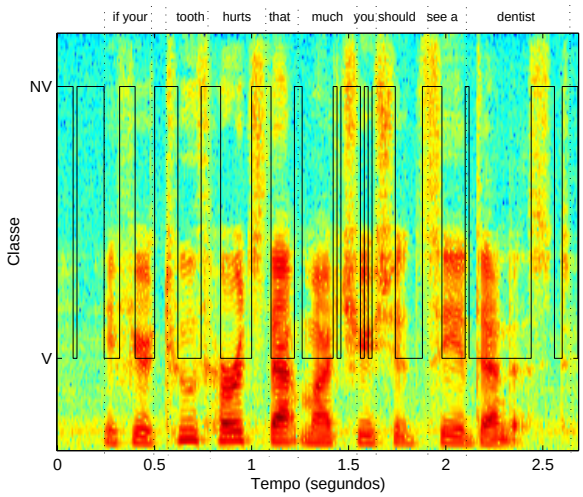


Figura 4.5: Classificação entre NV (não vozeado) e V (vozeado) para Arquivo Homem 2

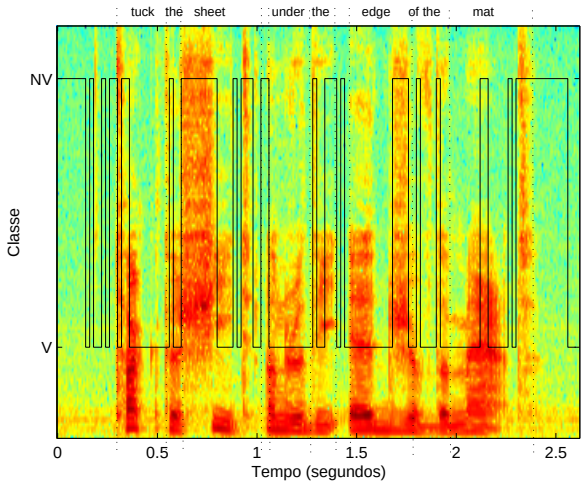
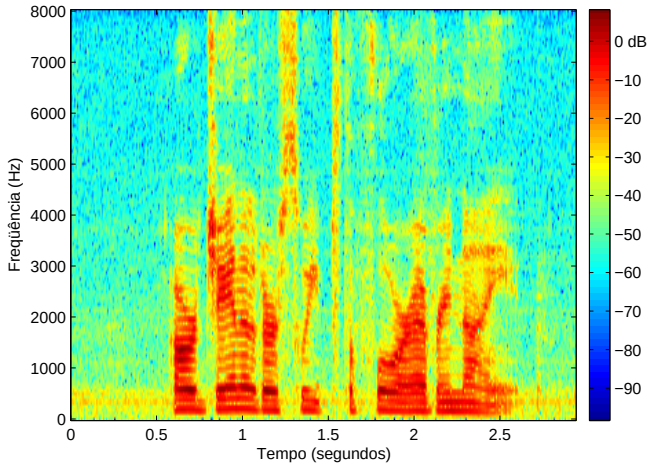
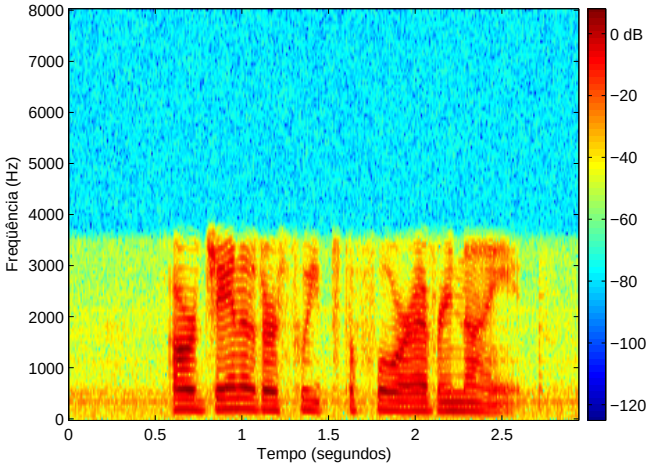


Figura 4.6: Classificação entre NV (não vozeado) e V (vozeado) para Arquivo Homem 3

neural obteve uma boa taxa de acerto em relação a identificar trechos não vozeados. Entretanto, a rede muitas vezes classifica equivocadamente trechos de silêncio como vozeados. Durante estes momentos a rede apresentou um comportamento anômalo, ora classificando como vozeado ora como não vozeado.

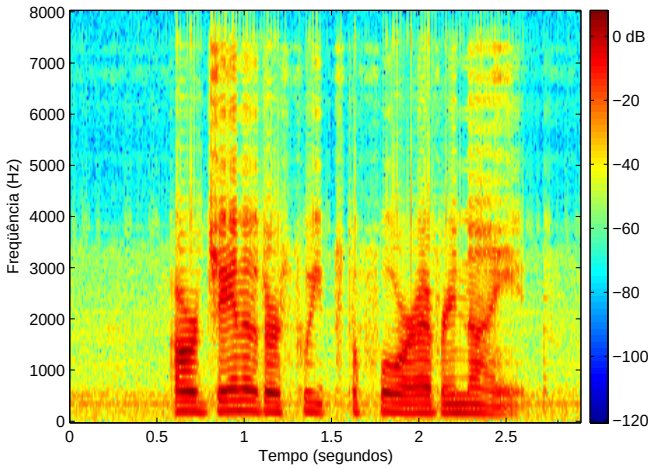


(a) Sinal de banda estreita $s_{tel16k}(n)$

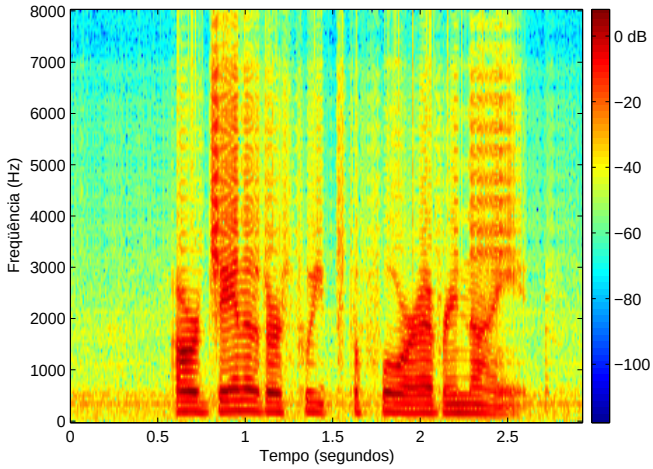


(b) Sinal de banda larga $s(n)$

Figura 4.7: Arquivo: Mulher 1



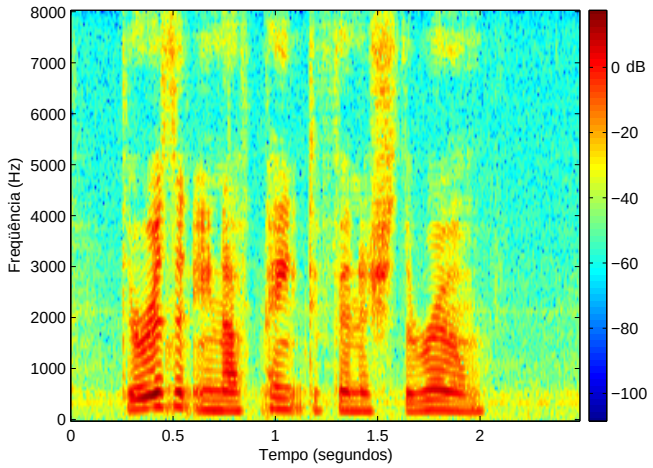
(a) Sinal gerado pelo algoritmo 1



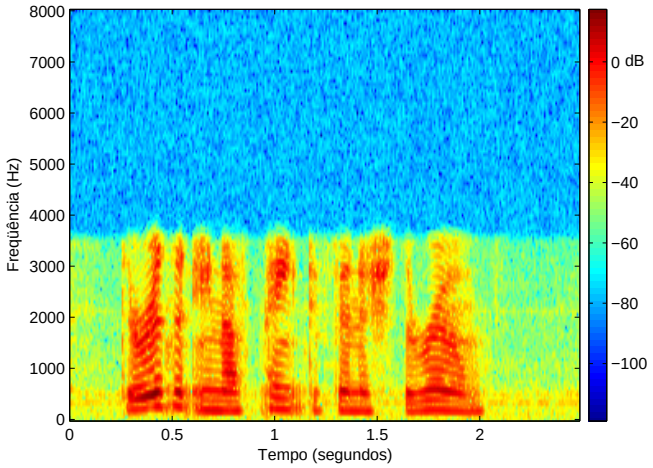
(b) Sinal gerado pelo algoritmo 2

Figura 4.8: Resultados Arquivo: Mulher 1

Observando as figuras que apresentam o sinal estendido pelo Algoritmo 1 é visto que em geral trechos classificados como vozeados são menos estendidos que os não vozeados. Porém para frases de locutores masculinos os baixos ganhos encontrados para estender trechos vozeados são grandes o suficiente para gerar superestimação da potência de trechos classificados como vozeados, como por exemplo na Figura 4.14(a) no final da palavra *write* dita pelo Homem 1. Isso indica que devem ser encontrados ganhos diferentes para

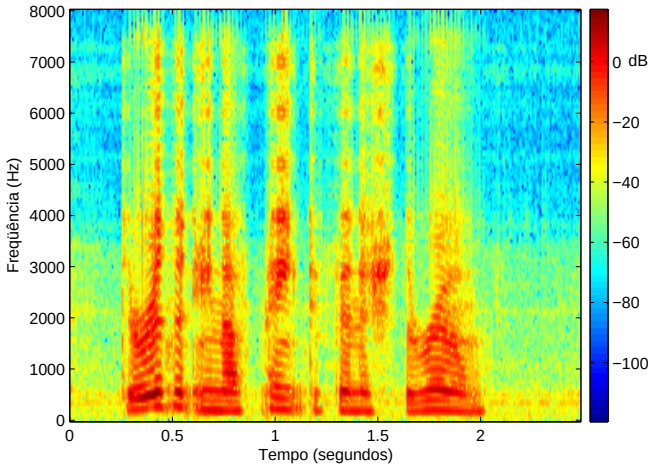


(a) Sinal de banda estreita $s_{tel16k}(n)$

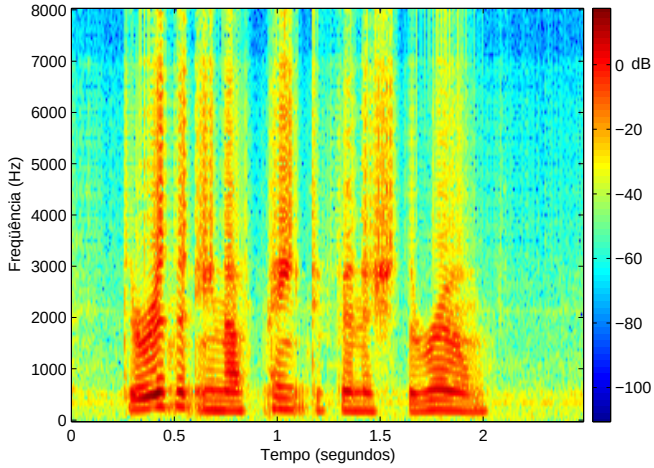


(b) Sinal de banda larga $s(n)$

Figura 4.9: Arquivo: Mulher 2

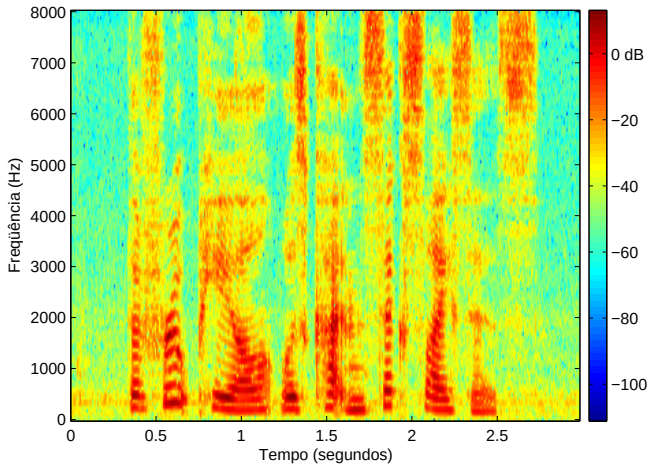


(a) Sinal gerado pelo algoritmo 1

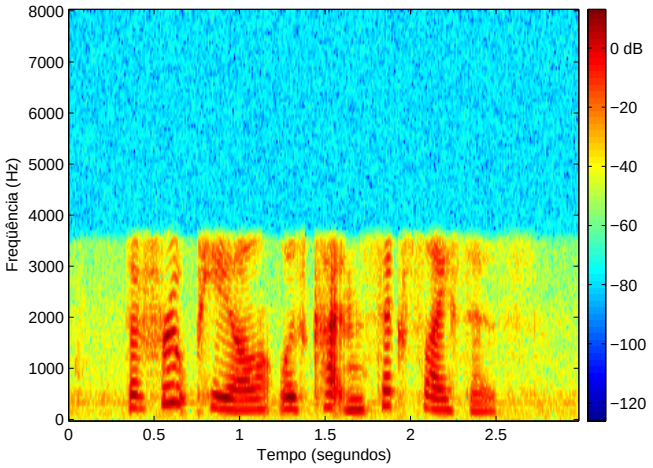


(b) Sinal gerado pelo algoritmo 2

Figura 4.10: Resultados Arquivo: Mulher 2

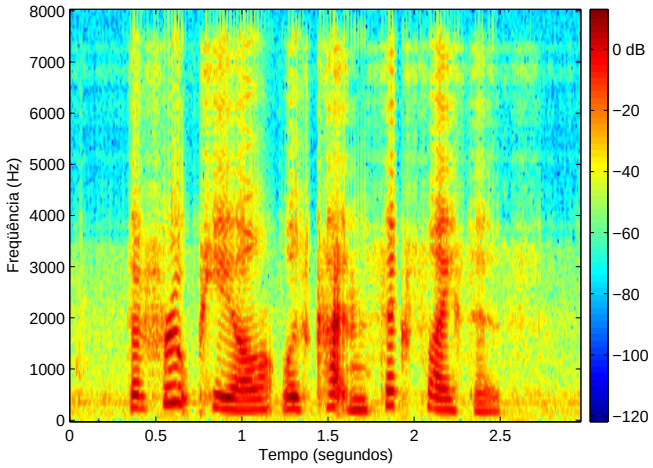


(a) Sinal de banda estreita $s_{tel16k}(n)$

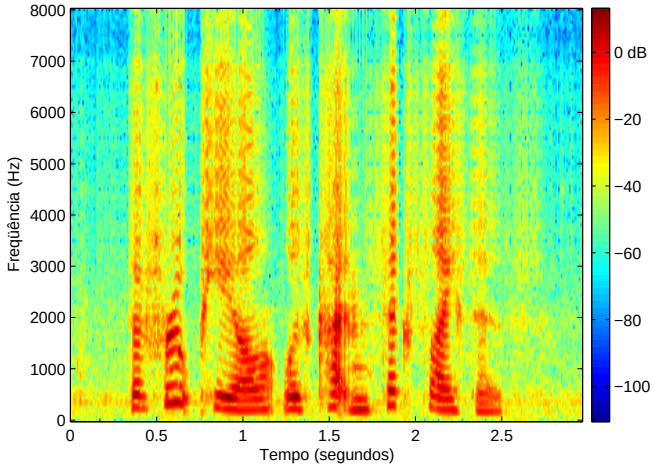


(b) Sinal de banda larga $s(n)$

Figura 4.11: Arquivo: Mulher 3

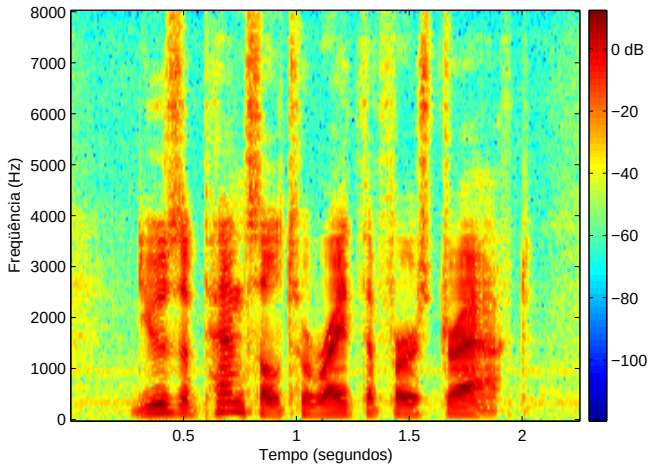


(a) Sinal gerado pelo algoritmo 1

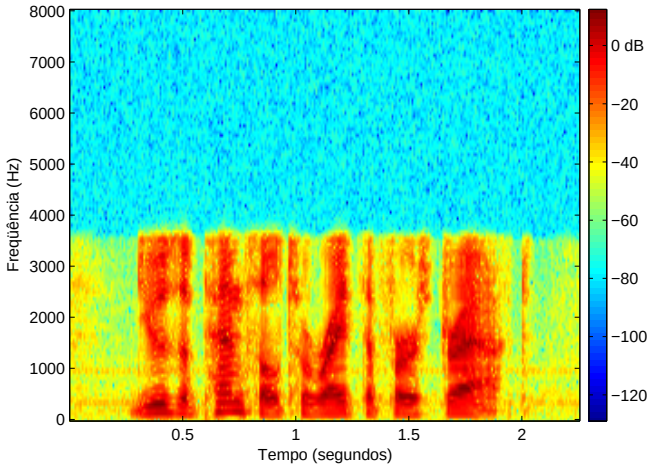


(b) Sinal gerado pelo algoritmo 2

Figura 4.12: Resultados Arquivo: Mulher 3

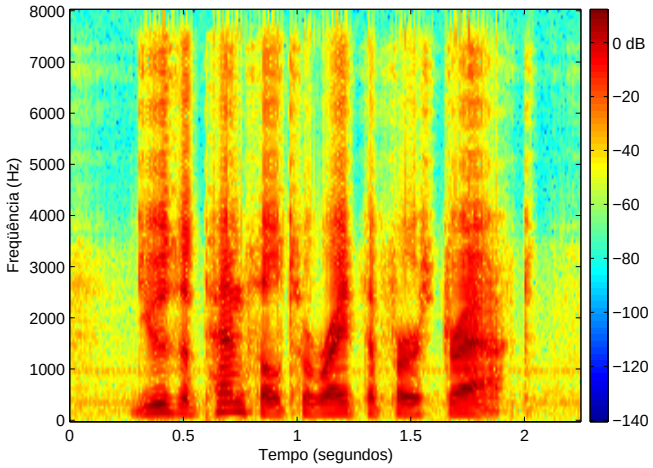


(a) Sinal de banda estreita $s_{tel16k}(n)$

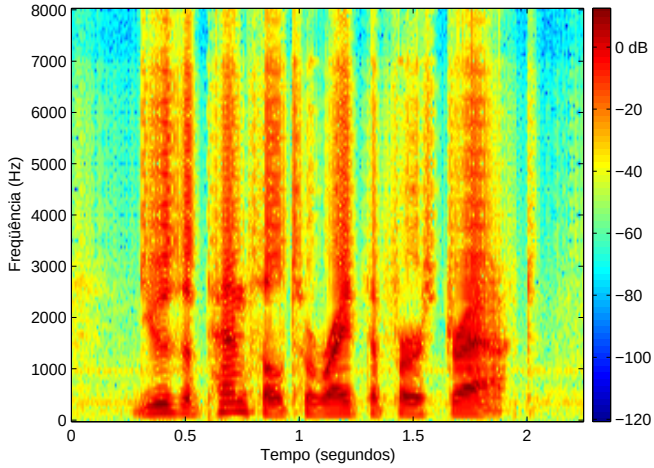


(b) Sinal de banda larga $s(n)$

Figura 4.13: Arquivo: Homem 1

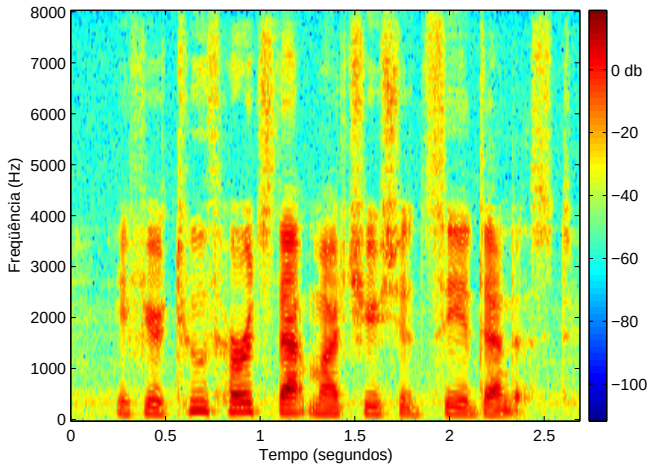


(a) Sinal gerado pelo algoritmo 1

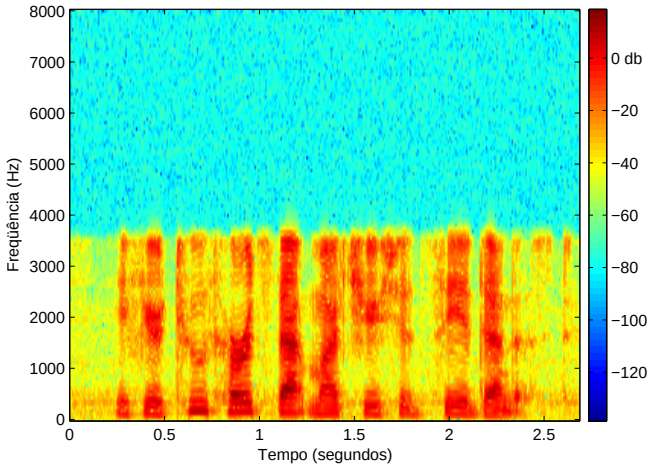


(b) Sinal gerado pelo algoritmo 2

Figura 4.14: Resultados Arquivo: Homem 1

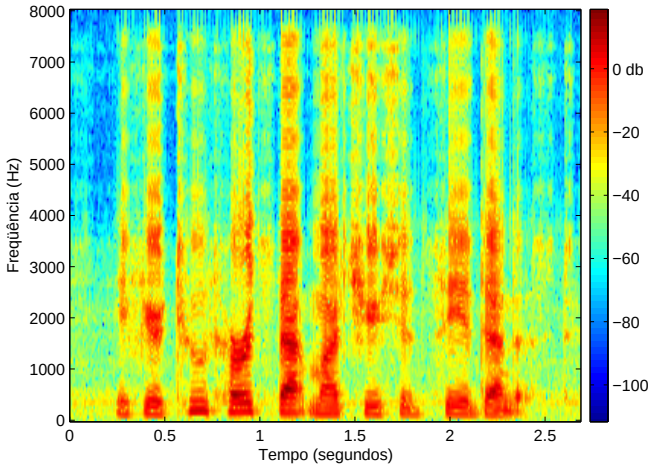


(a) Sinal de banda estreita $s_{tel16k}(n)$

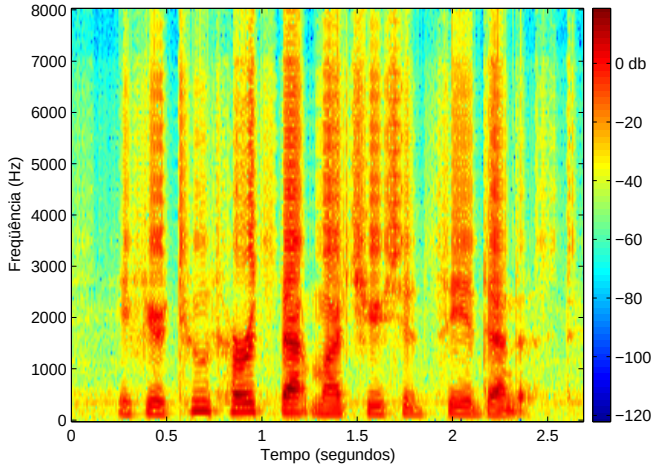


(b) Sinal de banda larga $s(n)$

Figura 4.15: Arquivo: Homem 2

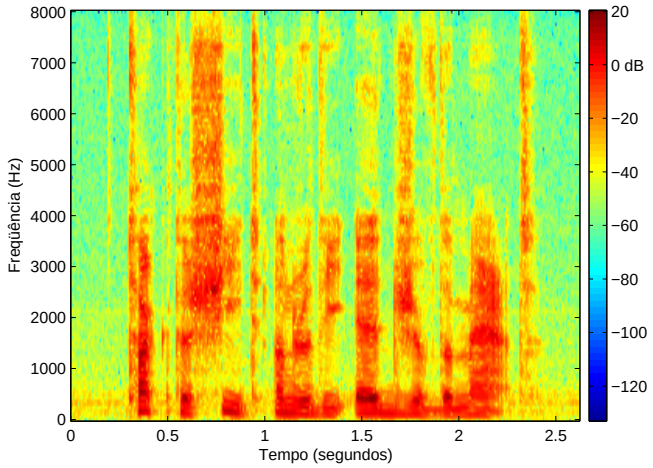


(a) Sinal gerado pelo algoritmo 1

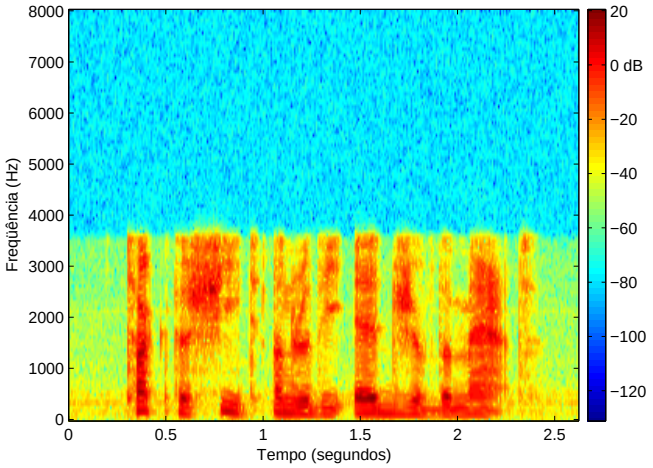


(b) Sinal gerado pelo algoritmo 2

Figura 4.16: Resultados Arquivo: Homem 2

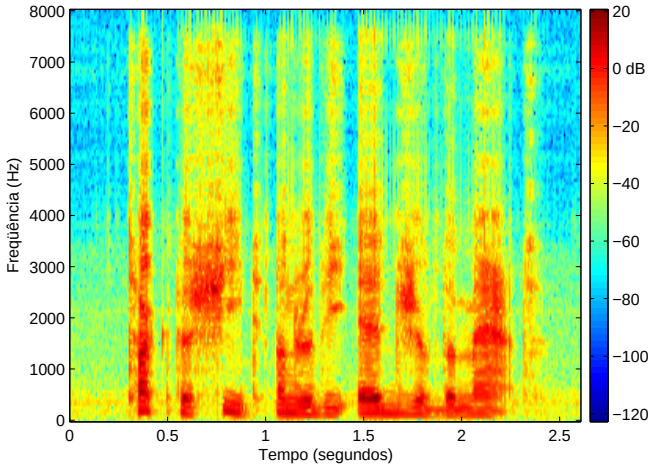


(a) Sinal de banda estreita $s_{tel16k}(n)$

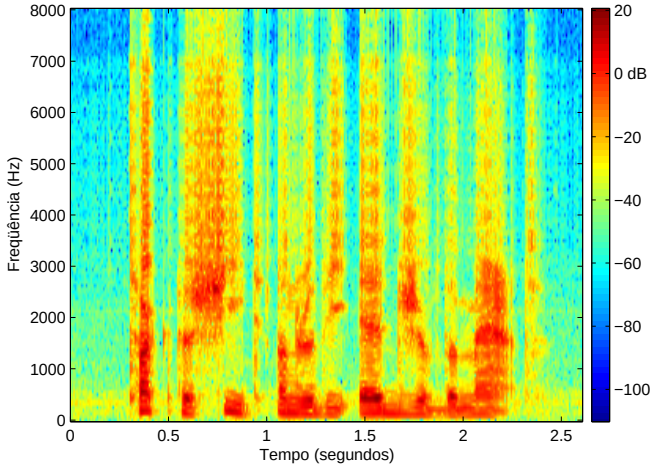


(b) Sinal de banda larga $s(n)$

Figura 4.17: Arquivo: Homem 3



(a) Sinal gerado pelo algoritmo 1



(b) Sinal gerado pelo algoritmo 2

Figura 4.18: Resultados Arquivo: Homem 3

homem e mulher. Pode-se tecer o mesmo comentário em relação os filtros encontrados para estender sinais não vozeados, que subestimaram, em alguns trechos de voz feminina, a potência real das bandas superiores. Isso pode ser visto por exemplo na extensão do começo e final da palavra *slices* da Mulher 3 vista na Figura 4.12(a).

4.2 Análises

Os métodos de avaliação são essenciais para determinarmos se a performance de um algoritmo é melhor que a de outro. Existem diversos métodos quantitativos para avaliar a performance de algoritmos de extensão em freqüência, como por exemplo medidas de distância entre o sinal gerado da extensão e o sinal original de banda larga. Esses métodos são elaborados levando em consideração a estrutura do "envelope espectral" que modela o trato vocal do sinal de banda larga assim como as características do sistema auditivo do ser humano. Mas de uma maneira geral os testes subjetivos ainda apontam para uma avaliação mais confiável.

No resultado de uma avaliação subjetiva é levada em consideração a opinião de um grupo de pessoas que ouve o sinal estendido e classifica a performance do algoritmo de acordo com alguns parâmetros. Para avaliar os dois algoritmos realizamos uma análise quantitativa dos sinais gerados, no qual foi calculado o parâmetro LSD (*log spectral distortion*) para ambos algoritmos e para um terceiro algoritmo, e um teste subjetivo que contou com a participação de 20 pessoas, no qual os três algoritmos foram avaliados.

O Algoritmo 3 realiza a extensão assim como o Algoritmo 2, se apropriando da estrutura do modelo do trato vocal mas mapeando linearmente os coeficientes do filtro do sinal de banda estreita e banda larga. Ele foi incluído a fim de avaliarmos os resultados dos algoritmos aqui propostos, comparando-os com resultados de um método que se mostrou satisfatório.

4.2.1 Medidas quantitativas

A medida quantitativa que foi utilizada para comparar a extensão realizada por cada algoritmo foi o valor RMS do LSD (*log espectral distortion*) [16]. Esse parâmetro considera a estrutura do "envelope espectral" que modela o trato vocal do sinal s_{mb}, que contém somente a informação perdida na transmissão (as freqüências perdidas):

$$|S_{mb}(e^{jw})|^2 = \frac{\sigma_{mb}}{|A_{mb}(e^{jw})|} \tag{4.1}$$

É feita uma análise LPC do sinal s_{mb} e do sinal de banda estreita s_{tel}. A primeira análise resulta nos coeficientes a_{mb} e no ganho escalar σ_{mb} da banda perdida, e a segunda resulta também em um fator de ganho σ_{tel} da banda limitada, ambos ilustrados na Figura 4.19. Essas grandezas serão utilizadas para o cálculo do LSD, assim definido:

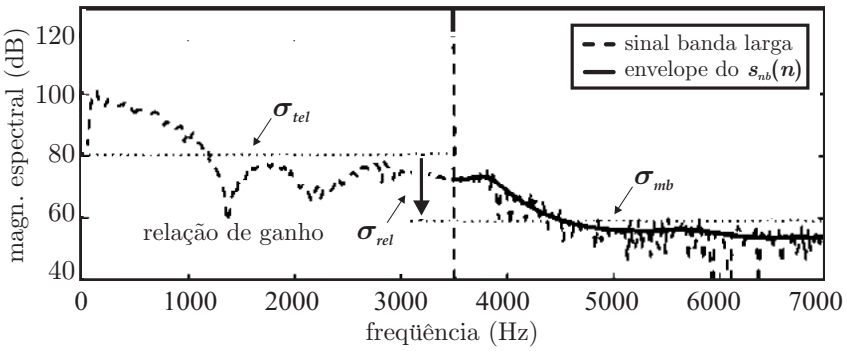


Figura 4.19: Relação das grandezas utilizadas para o cálculo do LSD

$$d^2_{\text{LSD}} = \frac{1}{2\pi}\int_{-\pi}^{\pi}(20\log_{10}\frac{\sigma_{rel}}{|A_{mb}(e^{jw})|} - 20\log_{10}\frac{\tilde{\sigma}_{rel}}{|\tilde{A}_{mb}(e^{jw})|})^2\ dw \tag{4.2}$$

no qual $\sigma_{rel} = \sigma_{mb}/\sigma_{bb}$ e as quantidades marcadas com " ˜ " são provenientes de cálculos realizados com os valores estimados na extensão em freqüência. A unidade do valor RMS d_{LSD} é dB.

O cálculo utilizado para obter o valor LSD foi:

$$d^2_{\text{LSD}} = \frac{1}{M}\sum_{K=0}^{M-1}(20\log_{10}\frac{\sigma_{rel}}{|A_{mb}(K)|} - 20\log_{10}\frac{\tilde{\sigma}_{rel}}{|\tilde{A}_{mb}(K)|})^2\ dw \tag{4.3}$$

no qual $A_{mb}(K)$ se refere ao K-ésimo termo da transformada discreta de Fourrier do filtro $A_{mb}(e^{jw})$ e M é o número de pontos da transformada.

Para cada sinal de voz testado foi medido o valor RMS do LSD em cada *frame* processado. As Tabelas 4.2, 4.3 e 4.4 contêm os valores da média e do desvio padrão obtidos para os Algoritmos 1, 2 e 3 para cada arquivo de voz.

d_{LSD} - RMS-LSD

	Mulher 1	Mulher 2	Mulher 3	Homem 1	Homem 2	Homem 3
média	9,6579	10,061	14,029	9,7197	10,142	10,108
desvio	5,9923	4,9616	8,1414	5,3275	5,8513	4,9638

Tabela 4.2: Valor médio e desvio do RMS-LSD calculado para vozes femininas e masculinas estendidas pelo Algoritmo 1

É importante atentar ao fato de que o LSD é uma medida de distorção , então quanto maior for seu valor maior será a distorção gerada pelo algoritmo em relação ao sinal de banda larga. As tabelas ilustram que para todas as frases testadas os menores valores médios e desvios do RMS do LSD são obtidos pelo Algoritmo 2, e os maiores pelo Algoritmo 1. Em geral os menores valores médios do RMS LSD são obtidos para frases com locutores femininos, exceto pelo Algoritmo 1, que obteve na média menores valores

d_{LSD} - RMS-LSD						
	Mulher 1	Mulher 2	Mulher 3	Homem 1	Homem 2	Homem 3
média	6,5247	6,2423	8,9055	8,0811	9,2914	7,0099
desvio	4,038	3,3618	5,3559	4,885	4,8664	3,804

Tabela 4.3: Valor médio e desvio do RMS-LSD calculado para vozes femininas e masculinas estendidas pelo Algoritmo 2

d_{LSD} - RMS-LSD						
	Mulher 1	Mulher 2	Mulher 3	Homem 1	Homem 2	Homem 3
média	8,6447	8,0713	9,8262	9,1665	10,244	8,4942
desvio	5,0612	4,3321	6,2276	4,8035	5,224	4,8934

Tabela 4.4: Valor médio e desvio do RMS-LSD calculado para vozes femininas e masculinas estendidas pelo Algoritmo 3

para frases ditas por homens.

4.2.2 Testes subjetivos

Foram realizados testes CMOS (*comparative mean opinion score)*, no qual a pessoa deveria comparar os sinais gerados por três algoritmos de extensão de freqüência diferentes com o sinal de banda estreita correspondente. Essa comparação seria realizada de acordo com dois parâmetros, um que levaria em consideração a qualidade da extensão em relação à presença ou não de ruídos de alta freqüência e artefatos, e o outro consideraria a extensão realizada pelo algoritmo, ou seja, quão próximo o resultado se assemelhou ao sinal de banda larga correspondente.

Essa diferenciação é importante pois estamos lidando com a possibilidade de um algoritmo implementar uma extensão cuja magnitude exceda o sinal original, introduzindo artefatos e ruídos no sinal, o que compromete a qualidade do som e incomoda o ouvinte. Para avaliar esses parâmetros, o ouvinte escolheu uma nota de uma escala de 5 níveis apresentada no formulário mostrado no Apêndice A.

Para cada uma das seis frases o participante ouve primeiramente o sinal de banda estreita e o sinal de banda larga original, para que perceba a perda que ocorre quando limitamos o sinal de voz na freqüência e para que seja capaz de avaliar o conceito de extensão.

Após ouvir esses dois sinais o teste prossegue com o sinal de banda estreita seguido pelo sinal de banda estendida gerado por um determinado algoritmo, tocado duas vezes. Esse esquema se repete para os outros dois algoritmos testados.

Dessa maneira o participante pode classificar os sinais criados pelos algoritmos comparando-os com o sinal de banda estreita para avaliar a extensão, e realizar uma comparação entre os sinais gerados artificialmente para avaliar a qualidade. Os algorit-

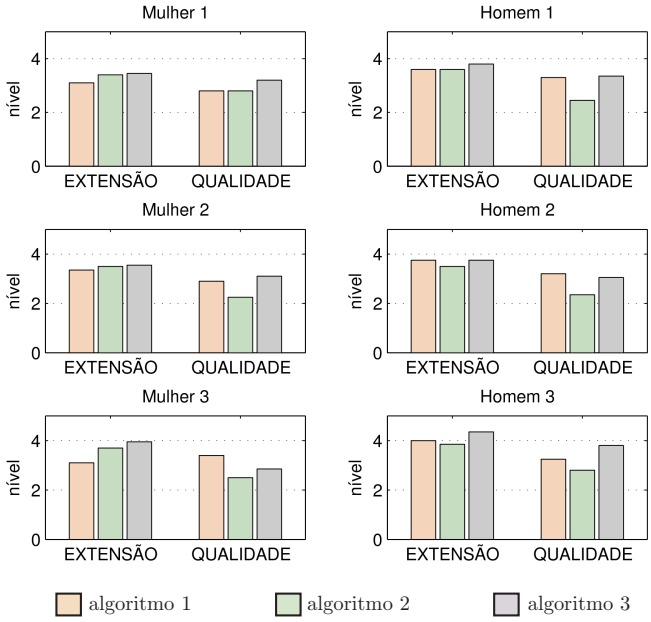


Figura 4.20: Níveis médios obtidos em cada frase para os três algoritmos

mos não foram apresentados na mesma ordem para frases diferentes a fim de não viciar as avaliações.

Ao total 20 testes foram realizados, com pessoas de idades e gêneros distintos. Os resultados expostos na Figura 4.20 foram gerados pela média das notas escolhidas pelos participantes para um determinado algoritmo para cada frase, em relação ao parâmetro extensão e qualidade.

É visto que nem sempre as análises quantitativas e subjetivas apontam para a mesma direção. Se para o parâmetro de distorção espectral o Algoritmo 2 obteve resultados claramente superiores, nos testes subjetivos é necessário uma análise mais detalhada para avaliar a performance dos algoritmos.

Pelos resultados expostos na Figura 4.20 nota-se que em geral o Algoritmo 2 estende mais que o Algoritmo 1 quando o locutor é uma mulher. Esse resultado se altera quando para locutores masculinos. Em relação à qualidade, o Algoritmo 1 apresentou níveis superiores ao Algoritmo 2 em todas frases testadas.

Esses resultados mostram que a extensão realizada pelo Algoritmo 2 foi em alguns casos superestimada, introduzindo no sinal de telefone informação que deteriorou a qualidade deste. Isso explica também os resultados que este algoritmo alcançou para locutores femininos, que devido à natureza da voz da mulher, exigem uma maior extensão, sendo esta característica atendida com maior êxito pelo Algoritmo 2.

Comparando os algoritmos propostos com o Algoritmo 3 nota-se que o último obteve níveis superiores aos dois primeiros em relação à extensão. Porém, no critério de qualidade o Algoritmo 1 alcançou níveis superiores para a frase Mulher 3 e Homem 2 e um nível

muito próximo na frase Homem 1.

Capítulo 5

Conclusões e Trabalhos Futuros

Os algoritmos de extensão em freqüência do sinal de telefone têm como objetivo melhorar a percepção auditiva do sinal que chega na linha do assinante. Essa melhoria não é feita no sentido de retirar possíveis ruídos introduzidos pelo canal de transmissão, mas sim, em relação à perda que ocorre quando o sinal de voz tem sua banda de freqüência limitada para atender as taxas atuais de transmissão. Essa perda não compromete o entendimento do sinal recebido no aparelho de telefone, mas introduz um efeito de "abafamento" no sinal original de banda larga. O assinante do sistema de telefonia atualmente, acostumado com a qualidade do áudio oferecida nas mídias mais recentes, é cada vez mais sensível aos efeitos dessa perda. Vista a necessidade de atenuá-la, considera-se válida a proposta deste trabalho em introduzir algoritmos capazes de gerar sinais, que segundo testes subjetivos, foram capazes de se aproximar do sinal originalmente transmitido.

O algoritmo de extensão em freqüência seria uma melhoria a mais introduzida no próprio aparelho do assinante, que poderia "ligar"ou "desligar"essa opção, de acordo com o que lhe convir. Foi importante, portanto, realizar testes que medissem não somente a capacidade desses algoritmos de atenuar as perdas impostas pela limitação da banda, mas também avaliar a qualidade do sinal gerado em relação a presença de artefatos e ruídos. Isto porque um assinante talvez prefira escutar o sinal abafado recebido ao sinal gerado pelo algoritmo de extensão, caso este possua ruídos que o incomodem.

Neste trabalho foram apresentados dois algoritmos de extensão em freqüência do sinal de telefone, um que realiza a extensão através de filtros e ganhos de acordo com a classificação entre sinais vozeados e não vozeados, e o outro que realiza a extensão através da obtenção de um modelo do trato vocal encontrado para cada subbanda do sinal.

Testes subjetivos indicam que o primeiro algoritmo apresenta em geral uma performance superior em relação a não inserção de ruídos e artefatos, e o Algoritmo 2 apresenta melhores resultados em relação à proximidade ao sinal de banda larga quando as frases são pronunciadas por locutores femininos.

Como trabalhos futuros é certo discernir algumas questões que merecem maior estudo. Para evitar que ocorra extensão durante períodos de silêncio seria interessante

treinar as redes de ambos algoritmos para serem capazes de identificar quando um quadro de silêncio é processado. Isso exigiria um estudo para precisar quais seriam os parâmetros de uma voz limitada na freqüência mais relevantes na identificação de períodos de silêncio.

É possível melhorar os resultados encontrados com a extensão gerada pelo Algoritmo 1 diferenciando vozes femininas e masculinas, ou seja, aplicando ganhos e filtros diferentes para cada caso. Seria necessário, neste caso, a identificação prévia do gênero do locutor, o que poderia ser realizado pela mesma rede neural utilizada para classificar vozeado e não vozeado, dependendo da relevância dos seus parâmetros de entrada na diferenciação do gênero do locutor.

Estuda-se também a importância da realização de testes subjetivos com frases em português, a língua nativa dos voluntários que participaram dos testes, uma vez que possibilitaria uma interpretação mais rica dos sinais gerados.

Bibliografia

[1] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2004, vol. 1, pp. 713–716.

[2] M. Nilsson, S.V. Andersen, and W.B. Kleijn, "On the mutual information between frequency bands in speech," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2000, vol. 3, pp. 1327–1330.

[3] U. Kornagel, "Spectral widening of the excitation signal for telephone-band speech enhancement," in *Proc. Int. Workshop Acoustic Echo and Noise Control*, 2001, pp. 215–218.

[4] Bernd Iser and Gerhard Schimidt, "Neural network versus codebooks in an application for bandwidth extension of speech signal," in *Proc. 8th European Conf. Speech, Commun. Tech.*, 2003, pp. 565–568.

[5] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluitjer, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2001, vol. 1, pp. 665–668.

[6] S. Haykin, *Modern Filters*, Prentice-Hall, 4th. edition, 2001.

[7] Harald Gustafsson, Ulf A. Lindgren, and Ingvar Claesson, "Low-complexity feature-mapped speech bandwidth extension," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 577–588, 2006.

[8] S. L. Marple Jr., *Digital Spectral Analysis with Applications*, Prentice-Hall, 4th. edition, 1987.

[9] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, 1993.

[10] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, 1998.

[11] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 2001.

[12] T. Q. Nguyen, "Near perfect reconstruction pseudo-qmf banks," *IEEE Trans. Signal Processing*, vol. 42, no. 1, pp. 65–76, 1994.

[13] G. Stang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley - Cambridge Press, 1996.

[14] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2004, vol. 1, pp. 697–700.

[15] R. O. Duda, P. E. Hart, and D. H. Stork, *Pattern Classification*, Wiley Interscience, 2nd. edition, 2000.

[16] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2002, vol. 1, pp. 237–240.

[17] G. Miet, A. Gerrits, and J. C. Valiere, "Low-band extension of telephone-band speech," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2000, vol. 3, pp. 1851–1854.

[18] D.A. Heide and G.S. Kang, "Speech enhancement for bandlimited speech," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1998, vol. 1, pp. 3931–396.

[19] Bernd Iser and Gerhard Schimidt, "Bandwidth extension of telephony speech," *Short Tutorials*, vol. 1, pp. 1–24, 2005.

[20] Philip D. Wasserman, *Neural Computing - Theory and Practice*, Van Nostrand Reinhold, 1989.

Apêndice A

Formulário de Avaliação dos Testes Subjetivos

1) Mulher 1:

	Extensão					**Qualidade**				
	1	2	3	4	5	1	2	3	4	5
I										
II										
III										

2) Homem 1:

	Extensão					**Qualidade**				
	1	2	3	4	5	1	2	3	4	5
I										
II										
III										

3) Mulher 2:

	Extensão					**Qualidade**				
	1	2	3	4	5	1	2	3	4	5
I										
II										
III										

4) Homem 2:

	Extensão					**Qualidade**				
	1	2	3	4	5	1	2	3	4	5
I										
II										
III										

5) Mulher 3:

	Extensão					**Qualidade**				
	1	2	3	4	5	1	2	3	4	5
I										
II										
III										

6) Homem 3:

	Extensão					**Qualidade**				
	1	2	3	4	5	1	2	3	4	5
I										
II										
III										