UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA ELÉTRICA
PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ANÁLISE E SÍNTESE DA FALA POR INTERPOLAÇÃO DE ONDAS

ACELINO DE CARVALHO COSTA FILHO

FEVEREIRO
2005

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA ELÉTRICA
PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ANÁLISE E SÍNTESE DA FALA POR INTERPOLAÇÃO DE ONDAS

Tese apresentada por **Acelino de Carvalho Costa Filho** à Universidade Federal de Uberlândia como parte dos requisitos para a **obtenção do título de Doutor em Engenharia Elétrica**, aprovada em 11/02/2005 pela Banca Examinadora:

Professor: **Gilberto Arantes Carrijo,** Dr. (UFU), Orientador

Professor: **Miguel Arjona Ramírez,** Dr. (USP)

Professor: **Sibelius Lellis Vieira,** Dr. (UCG)

Professora: **Edna Lúcia Flôres,** Dra. (UFU)

Professor: **Antônio Cláudio P. Veiga,** Dr. (UFU)

FICHA CATALOGRÁFICA

Elaborada pelo Sistema de Bibliotecas da UFU / Setor de Catalogação e Classificação

C837a Costa Filho, Acelino de Carvalho, 1957-
Análise e síntese da fala por interpolação de ondas / Acelino de Carvalho Costa Filho. - Uberlândia, 2005.
287f. : il.
Orientador: Gilberto Arantes Carrijo.
Tese (doutorado) – Universidade Federal de Uberlândia, Programa de Pós-Graduação em Engenharia Elétrica.
Inclui bibliografia.
1. Processamento de sinais - Teses. 2. Codificador de voz - Teses. I. Carrijo, Gilberto Arantes. II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Engenharia Elétrica. III. Título.

CDU: 621.391 (043.3)

ANÁLISE E SÍNTESE DA FALA POR INTERPOLAÇÃO DE ONDAS

ACELINO DE CARVALHO COSTA FILHO

Tese apresentada por Acelino de Carvalho Costa Filho à Universidade Federal de Uberlândia, como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica.

Prof. Gilberto Arantes Carrijo, Dr. Orientador	Prof. Gilberto Arantes Carrijo, Dr. Coordenador do Curso de Pós-Graduação

Aos meus pais,

Acelino e Maria Eunice,

À minha esposa,

Eufrosina,

E às minhas filhas,

Sofia e Olivia.

AGRADECIMENTOS

Ao Prof. Dr. Gilberto Arantes Carrijo, pela orientação e pela confiança depositada;

Ao Prof. Dr. Miguel Arjona Ramírez pela colaboração e sugestões apresentadas;

Aos professores da banca examinadora por aceitarem o convite e pelas discussões e correções que muito valorizam este trabalho;

Ao CNPq, pelo apoio financeiro parcial;

Ao CEFET – Go e Universidade Católica de Goiás pela oportunidade e incentivo;

À pequena Júlia por emprestar sua voz infantil;

Aos colegas do curso de doutorado e mestrado, em especial aos colegas do CEFET – Go e da UCG, pela convivência e constante espírito de solidariedade durante esta árdua jornada na UFU – Uberlândia;

À minha família que tanto se sacrificou para que este trabalho fosse realizado, aos familiares pelo constante incentivo e sobretudo à minha mãe que não viu a conclusão deste trabalho;

Aos professores e funcionários da Faculdade e da Pós – Graduação em Engenharia Elétrica, em especial ao professor Paulo Sérgio Caparelli e à secretária Joana Proença pela colaboração;

Aos colegas do MAF - UCG pelo apoio e incentivo, em especial aos profs. Newton Olímpio e Juan Barrio pelo constante incentivo, credibilidade e empenho em resolver reveses com justiça, sabedoria e perspicácia;

Aos colegas no CEFET – Go, em especial a equipe da física pelo companheirismo, compreensão, apoio e incentivo e ao prof. Jorge Antônio de Souza pela amizade e pela colaboração nas horas mais difíceis;

E a todos aqueles que participaram direta ou indiretamente colaborando para que este trabalho fosse realizado.

RESUMO

Costa Filho, A. C. Análise e Síntese da Fala por Interpolação de Ondas, Uberlândia, Digra-UFU, 2005, 287 pp.

Esta tese aborda o estudo da técnica de codificação da fala por interpolação de ondas, técnica WI, e a descrição, visualização e implementação de um algoritmo para simular um sistema de análise - síntese WI convencional do sinal da fala, como parte de um *sistema de codificação – decodificação WI que opera na camada externa, sendo então denominado de sistema de análise - síntese WI (padrão)*. A partir desse sistema são efetuadas pesquisas buscando melhorias no processo de localização e extração das formas de ondas características, as CW's, que são obtidas por interpolação nas posições regulares durante o estágio de análise. São apresentadas propostas, sendo que parte delas foram implementadas resultando no sistema denominado de *sistema de análise - síntese WI (com interpolação)*. Os resultados das simulações e as avaliações que mostram o desempenho do *sistema de análise – síntese WI (padrão)* e do *sistema de análise - síntese WI (com interpolação)* foram apresentados, para alguns sinais da fala. Na avaliação destes sistemas WI foram utilizados os métodos PESQ, para verificar a qualidade perceptual da fala, e o método SNRSEG - NCCF desenvolvido neste trabalho, para verificar a reconstrução das formas de ondas e a defasagem por segmento dos sinais originais e reconstruídos. O *sistema de análise – síntese WI (padrão)* apresentou boa reconstrução do sinal indicado pela similaridade entre as envoltórias temporais e entre as formas de ondas, e pelas medidas SNRSEG-NCCF. O sistema WI (padrão) também apresentou uma defasagem menor que o sistema de análise – síntese WI (com interpolação) e qualidade perceptual da fala reconstruída entre "satisfatória" e "boa" situada na faixa "*communication quality*" conforme a classificação MOS. A versão preliminar do sistema de análise – síntese WI (com interpolação) apresentou melhor *reconstrução das formas de ondas por segmento* e para alguns trechos dos sinais uma melhor qualidade perceptual da fala que o sistema de análise – síntese WI (padrão), mas com um grau de *defasagem por segmento maior* entre os sinais originais e reconstruídos. Em geral a qualidade da fala reconstruída no sistema WI (com interpolação) foi inferior ao sistema WI (padrão), mas os resultados indicam que o modelo tem potencial para apresentar resultados superiores com ajuste no processo de interpolação que melhore a sincronia dos sinais.

Palavras-chave: *Processamento da Fala, Codificação da Fala, Codificação da Fala por Interpolação de Ondas, Técnica WI, Análise e Síntese da Fala por Interpolação de Ondas.*

ABSTRACT

Costa Filho, A. C. Análise e Síntese da Fala por Interpolação de Ondas, Uberlândia, Digra-UFU, 2005, 287 pp.

This doctoral dissertation addresses to study of waveform interpolation (WI) for speech coding, and the description, visualization and implementation of an algorithm to simulate a conventional WI system of analysis – synthesis of speech signal, like portion of a WI system for coding - decoding that operates on the outer layer which is called of WI analysis synthesis system (standard). From this system researches are performed looking for improvement on the localization and extraction process of characteristic waveforms (CWs), that are obtained through interpolation at regular positions at the analysis stage. The propositions for improvement are presented, where part of them were implemented resulting on the system called the WI analysis – synthesis system (with interpolation). The results of simulations and assessments, that indicate the performance of WI analysis – synthesis system (standard) and WI analysis – synthesis system (with interpolation) were presented with some speech signals. For assessment of the WI systems, the PESQ method were utilized to verify the perceptual quality of speech, and the SNRSEG-NCCF method, developed on this work, to verify the reconstruction of the waveforms and the phase delay per segment between the original and reconstructed speech signals. The WI analysis – synthesis system (standard) presented good reconstruction of signal shown by the similarity between the time envelopes and the waveforms, and for the SNRSEG-NCCF measures. The WI system (standard) also presented a smaller phase displacement than the WI analysis – synthesis system (with interpolation) and the perceptual quality of reconstructed speech ranged between "satisfactory" and "good" located on the band "communication quality" according the MOS rating. The preliminary version of the WI analysis – synthesis system (with interpolation) presented better waveform reconstruction by segment and, for some parts of the signals, a better perceptual quality of speech than the WI analysis – synthesis system's (standard), but with a larger phase displacement by segment between the originals and the reconstructed speech signals. In general the quality of reconstructed speech at the WI system (with interpolation) was lower than the WI system (standard), but the results indicate that the model has potential to present enhanced results by adjusting the interpolation process to improve the synchrony of signals.

Keywords: *speech processing, speech coding, waveform interpolation, waveform interpolation speech coding, WI speech coding, WI technique, WI analysis – sinthesis of speech.*

ANÁLISE E SÍNTESE DA FALA POR INTERPOLAÇÃO DE ONDAS

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DAS PRINCIPAIS ABREVIATURAS

AbS	"*Analysis-by-Synthesis*"
ADM	Codificação por modulação delta adaptativa
APC	Codificação Preditiva adaptativa
APCM	Codificação por modulação de pulso adaptativo
ATC	"Adaptive Transfom Coding"
BSD	"*Bark Spectral Distance*"
CCITT	"*International Telegraph and Telephone Consultative Committee's*"
CELP	"*Code – Excited Linear Preditive Coders*"
CW	"*characteristic waveform*"
DoD	"*Department of defense (USA)*"
DRT	"*Diagnostic Rhyme Test*"
DAM	"*Diagnostic Acceptability Measure*"
DCT	"Discrete Cosine Transform"
DFT	"Discrete Fourier Transform"
DM	Codificação por modulação delta
DPCM	"*Differential Pulse Code Modulation"* - Codificação por modulação de pulso diferencial
DTFS	"*Discrete-Time Fourier Series*"
FACN	função autocorrelação normalizada
FCCN	função correlação cruzada normalizada
GSM	"*Global System Mobile*"
IMBE	"*Improved Multi-Band Excitation coder*"
ITU-T	"*Telecommunication Standardization Sector of International Telecommunication Union*"
LPC	"*Linear Predictive Coding*"
LPC	"*Linear Predictive Speech Coding*"
LP	"*Linear Predictive*"
LP	*"Linear Prediction Analysis"*
LP	*"Linear Prediction Synthesis"*
LPAS	*Codificadores de Análise por Síntese Baseados em Predição Linear*
LSF	"*Line Spectral Frequency*"
LSP	"*Line Spectral Pairs*"
MBE	"*Multi-Band Excitation"*
MELP	"*Mixed Excitation Linear Prediction*"

MPE	"*Multi-Pulse Excited Coder"*
MOS	"*Mean Opinion Score*"
PCM	"*Pulse Code Modulation*"
PESQ	"*Perceptual Evaluation of Speech Quality*"
PSQM	"*Perceptual Speech Quality Measure*"
PWI	"*Prototype Waveform Interpolation*"
REW	"*Rapidly evolving waveform*"
RMS	"*Root Mean Square log spectral distortion"*
RPE	*"Regular Pulse Excitated*"
SBC	"*Sub-band Coding*"
SEGSNR	"*Segmental SNR*" - *Relação sinal ruído segmental*
SEW	"*slowly evolving waveform*"
SNR	"*Signal-to-Noise Ratio*"
SNRSEG	Relação Sinal Ruído Segmental
SNRSEG-NCCF	medida obtida a partir do cálculo da Relação Sinal Ruído Segmental (SNRSEG) com o auxilio da "Normalized Cross Correlation Function" (NCCF)
STC	"*Sinusoidal Transform Coding*"
STFT	"*Short-Term Fourier Transform*"
TFI	"*Time Frequency Interpolation*"
TIA	"*Telecommunications Industry Association*"
VOCODERS	"*voice coders*"
VoIP	"*Voice Over Internet Protocol*"
WI	"*Waveform Interpolation (WI) Speech Coding*" - Técnica WI – Codificação da Fala por interpolação de ondas

Capítulo 1

INTRODUÇÃO

1.1 Considerações iniciais

A fala é o meio mais comum da comunicação humana sendo também o mais utilizado nas redes de telecomunicações. Ela distingue o ser humano dos outros seres vivos, permitindo trocar idéias, expressar opiniões ou revelar o pensamento de cada pessoa. No mundo que hoje é chamado de aldeia global, a rede telefônica e a Internet que suportam a transmissão de informação falada estão em constante evolução. Por outro lado, com o aparecimento dos sistemas automáticos existe também uma maior interação do homem com estes sistemas (interação homem – máquina) pelo desenvolvimento de interfaces que utilizam a fala. Assim, o homem passa a não ser o único gerador ou destinatário da fala, pois parte desta cadeia pode ser implementada por sistemas automáticos. Neste contexto, a fala que representa um componente central da comunicação digital, constitui a maior impulsionadora das tecnologias de telecomunicações. A representação dos sinais da fala com eficiência tem se tornado cada vez mais importante, onde a codificação da fala é um dos aspectos fundamentais.

O objetivo do processamento digital de sinais da fala é a representação digital destes sinais, a análise e a extração das características e o desenvolvimento dos modelos de síntese. Todas estas ferramentas são cruciais na implementação dos sistemas de comunicação falada, seja esta comunicação à distância ou a comunicação homem-máquina. Tradicionalmente, estes sistemas estão divididos em sistemas de codificação, síntese, reconhecimento de sinais da fala e verificação e identificação do locutor. Neste trabalho o interesse principal é a codificação da fala.

Este capítulo apresenta uma introdução à codificação da fala destacando sua importância nos sistemas de telecomunicações, a produção e a percepção da fala, a performance dos codificadores, os objetivos e o escopo desta pesquisa, e a organização desta tese.

1.2 A Codificação dos Sinais da Fala

Uma das mais antigas aplicações do processamento digital da fala é a codificação. Os codificadores foram utilizados primeiramente na criptografia dos sinais da fala como são utilizados hoje em sistemas de comunicação com segurança. Mas o principal uso da

codificação é na economia da taxa de bits para acomodar mais usuários em um canal de comunicações tais como células em telefonia móvel ou em um enlace de rede de pacotes [1]. Por codificação entende-se a representação eficiente do sinal (tentando diminuir a taxa de bits na representação do sinal) com vista à sua transmissão ou armazenamento, mas mantendo a qualidade acima da exigida pela aplicação. O esforço das pesquisas realizadas nesta área tem conduzido a diversas normas (recomendações) para trabalhar na rede telefônica pública, e mais recentemente tem sido feito esforços para utilizar também em redes baseadas em protocolos transmitindo pacotes de informação. No caso da transmissão de informação, a taxa de bits de codificação da fonte é um dos fatores mais importantes na definição da largura de faixa dos sistemas de comunicação. O armazenamento de grande quantidade de informação para a utilização posterior exige também a necessidade de reduzir a taxa de bits, já que ela determina o espaço requerido no meio do armazenamento. A necessidade de reduzir a taxa de bits permanece mesmo com o aumento da largura de faixa dos canais de transmissão, possibilitando transmitir um número maior de sinais no mesmo canal ou possibilitando utilizar canais ruidosos que precisam de códigos de detecção e correção de erros de grande redundância. Assiste-se ainda ao emergir dos serviços de multimídia com a integração de voz, imagens e dados, que necessitam racionalizar a distribuição da taxa de bits total por cada uma das aplicações.

Já há algum tempo que a codificação da fala tem sido facilitada pelo rápido avanço no processamento digital de sinais e pela capacidade dos processadores dos sinais digitais. Um incentivo forte para a pesquisa na codificação da fala é provido pelo deslocamento dos custos relativos envolvidos no manuseio da comunicação da fala nos sistemas de telecomunicações. Por outro lado, existe uma demanda crescente para aumentar a capacidades das redes de telecomunicações. Também, o rápido avanço na eficiência dos processadores de sinais digitais e das técnicas de processamento desses sinais tem estimulado o desenvolvimento de algoritmos de codificação da fala. Essas tendências provavelmente continuarão, e a codificação da fala permanecerá uma área de importância central para a redução dos custos de operação dos sistemas de comunicações de voz.

1.3 As Bases da Codificação da Fala

1.3.1 Propriedades do sinal da fala

Para construir um codificador da fala efetivo é necessário um bom conhecimento das propriedades do sinal da fala e de sua percepção. Tais conhecimentos levam aos modelos que removem as redundâncias do sinal da fala e que parametrizam somente as informações perceptuais relevantes. Enquanto o conhecimento existente das propriedades e da percepção

do sinal da fala é útil para o projeto de codificadores da fala, ele está longe de ser completo [2]. Na codificação da fala, a redução da taxa de bits é obtida pela remoção das redundâncias presentes na forma de onda do sinal.

1.3.1.1 Produção da fala

Atualmente é impossível construir um modelo do trato vocal e das cordas vocais que resulte em sons naturais da fala [2]. Quando a fala é produzida, um fluxo de ar é forçado a partir dos pulmões, passa pela laringe indo para o trato vocal. Na laringe, as cordas vocais elásticas podem obstruir o fluxo de ar, parcialmente ou totalmente, criando uma excitação vocal de ruído turbulento ou sopros de ar. A abertura entre as cordas vocais é denominada de glote, e o ar emanando a partir das cordas vocais é geralmente denominado de excitação glotal.

A fala pode ser, de forma geral, dividida em segmentos sonoros e surdos. Durante a fala sonora a glote abre e fecha periodicamente e a excitação glotal tem um caráter periódico. A forma de onda de excitação que corresponde a um ciclo da abertura e do fechamento glotal é referida como o pulso glotal, ou o pulso do *pitch*. Os pulsos do *pitch* consecutivos podem variar em seu comprimento e formato da onda, e a excitação glotal resultante é quase periódica.

Para a fala surda, a excitação é formada à medida que o ar passa pela glote, momento em que é forçado por uma constrição estreita em algum ponto do trato vocal criando uma turbulência. A glote não abre e fecha periodicamente, mas somente contrai causando perturbações no ar. As excitações surdas e segmentos surdos da fala não apresentam qualquer periodicidade aparente e têm um caráter ruidoso com menor energia quando comparado com o caso sonoro.

As propriedades temporais da produção da fala são refletidas nas características espectrais do sinal da fala. O espectro da excitação sonora tem uma estrutura harmônica, isto é, picos agudos em intervalos regulares de freqüência, com a freqüência fundamental correspondendo à taxa de fechamento da glote. O espectro da excitação surda não tem freqüências harmônicas proeminentes e assemelha-se ao espectro de um sinal de ruído branco. A excitação glotal não tem um envelope espectral característico exceto pela inclinação durante a fala sonora. O envelope espectral (os picos e os vales mais amplos no espectro) é imposto sobre a excitação glotal pelo trato vocal que atua como um filtro modelador da resposta em freqüência do sinal da fala.

A resposta em freqüência não plana do trato vocal introduz correlação de curto prazo entre as amostras adjacentes do sinal da fala. Durante a fala sonora o caráter periódico da excitação resulta na correlação entre as amostras correspondentes dos pulsos do *pitch* adjacentes (correlação de longo prazo). No domínio espectral, as correlações de curto prazo

correspondem ao envelope espectral e as correlações de longo prazo são refletidas na estrutura fina do espectro. Ambas as correlações introduzem informações redundantes no sinal da fala e podem ser exploradas na codificação da fala.

1.3.1.2 O sistema de audição humano

A percepção da fala é um processo complexo. Ainda não está claro como o sistema auditivo humano processa o sinal da fala [2]. Uma das propriedades usadas mais freqüentemente do sistema de audição humano é o fenômeno do mascaramento espectral. O mascaramento espectral faz com que a imprecisão na representação do sinal, devido à compressão, que ocorre nas faixas de freqüência com alta energia ou próximo delas seja menos audível do que as imprecisões que ocorrem em outras regiões. No domínio do tempo, o ouvido humano tem uma grande tolerância aos erros resultantes da imprecisão na representação das amostras de alta energia do que dos erros que coincidem com as amostras de baixa energia. Assim é evidente que ambas as características temporais e espectrais são importantes e isto está cada vez mais sendo utilizados nos codificadores modernos. De fato, os codificadores que combinam as análises no domínio da freqüência e no domínio do tempo são fortes concorrentes na área de codificação da fala em taxas muito baixas.

1.3.2 Atributos dos codificadores da fala

Um algoritmo de codificação da fala é avaliado com base nas seguintes características [2]: (i) a largura de faixa do sinal da fala que se pretende no codificador; (ii) a taxa de bits do sinal codificado; (iii) a qualidade do sinal reconstruído; (iv) o atraso devido ao algoritmo; (v) a complexidade do cálculo; (vi) a sensibilidade ao erro do canal; (vii) a robustez contra ruídos acústicos de fundo. Aplicações diferentes requerem que os codificadores sejam otimizados para diferentes características ou algum balanço entre elas. Por exemplo, em sistemas de transmissão de mensagens, o algoritmo de pequeno atraso pode não ser uma importante característica para o codificador, enquanto os sistemas de armazenamento central podem não requerer codificadores com implementação de pouca complexidade computacional. Para um grande número de aplicações o principal objetivo é assegurar a similaridade entre os sinais original e reconstruído, e em alguns casos, como nos sistemas em que a segurança é o principal objetivo, é suficiente que os sons da fala reconstruída sejam inteligíveis e naturais. Em geral, o compromisso principal da codificação é entre a taxa de bits do sinal codificado e a qualidade perceptiva da fala reconstruída. Na maioria das aplicações comerciais são

requeridas implementações em tempo real do codificador. Uma implementação em tempo real impõe restrições na complexidade computacional e no atraso do algoritmo do codificador.

1.3.3 Medidas da qualidade da fala

Uma das maiores dificuldades no projeto e no teste de um codificador é medir a qualidade da fala reconstruída. Esta dificuldade está principalmente na falta de uma medida objetiva para a qualidade da fala que represente a percepção em forma de uma função erro entre o sinal original e o sinal reconstruído. Existem duas maneiras típicas de medir a qualidade da fala: pelas medidas subjetivas da qualidade da fala e pelas medidas objetivas da qualidade da fala. Entretanto na tentativa de diminuir as dificuldades, mais recentemente, têm sido pesquisados testes automáticos e medidas objetivas que sejam capazes de predizer a qualidade subjetiva da fala [3, 4, 5, 6].

1.3.3.1 Medidas subjetivas da qualidade da fala

O procedimento de uma avaliação subjetiva é feito usualmente pelos testes de audição com respostas a um conjunto de sílabas, palavras, frases ou com outras questões. O material de teste é usualmente focalizado nas consoantes, por que elas são mais difíceis de serem sintetizadas do que as vogais. Nestes testes a qualidade da fala é usualmente medida pela inteligibilidade tipicamente definida como a porcentagem de palavras ou fonemas ouvidos corretamente, e pela naturalidade. Os três tipos de testes mais comuns usados para medir a qualidade da fala são:

- Teste do Diagnóstico de Rima ou ("*Diagnostic Rhyme Test*" (DRT)): O DRT é uma medida de inteligibilidade onde a tarefa do sujeito é reconhecer uma entre duas possíveis palavras em um conjunto de pares de rimas [7].

- Medida do Diagnóstico da Aceitabilidade ("*Diagnostic Acceptability Measure*" (DAM)): Os valores do DAM avaliam a qualidade do sistema de comunicação baseado na aceitabilidade da fala percebida por um ouvinte profissional treinado.
- Medida da Opinião Média ("*Mean Opinion Score*" (MOS)): A MOS é uma medida que é amplamente usada para quantificar a qualidade da fala reconstruída. A MOS usualmente envolve 12 a 24 ouvintes [8] (testes formais do Comitê Consultivo Internacional de Telegrafia

e Telefonia (CCITT)[1.1] e da Associação das Indústrias de Telecomunicações (TIA)[1.2] tipicamente envolvem de 32 a 64 ouvintes) que são instruídos para avaliar as gravações do sinal da fala balanceadas foneticamente em um dos 5 níveis de uma escala de qualidade. A escala de avaliação e sua descrição são apresentadas na Tabela 1.1 e na Figura 1.1. Observa-se que as medidas MOS podem diferir de teste para teste e portanto não são medidas absolutas para a comparação de diferentes codificadores.

Tabela 1.1 – Descrição do índice MOS relacionado com a qualidade da fala.

Avaliação	Qualidade da Fala	Nível de Distorção
5	Excelente	Imperceptível
4	Boa	Apenas perceptível, mas não irritante
3	Satisfatório	Perceptível e levemente irritante
2	Razoável	Irritante, mas não desagradável
1	Ruim	Muito irritante e "objectionable"

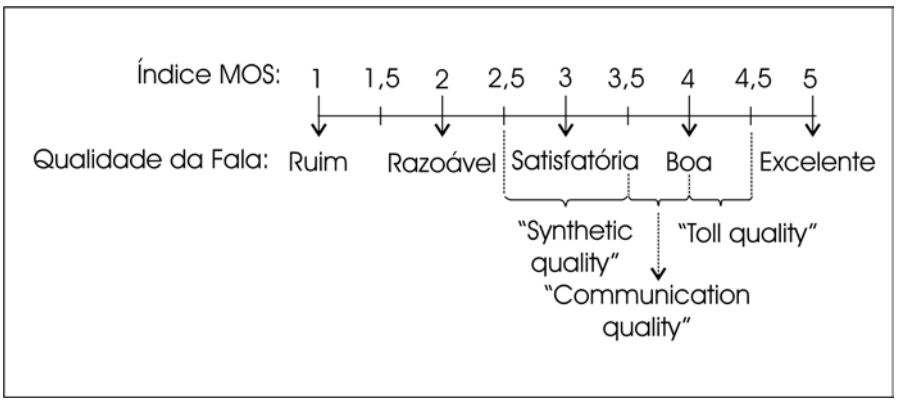

Figura 1.1 – Representação esquemática do índice MOS e da qualidade da fala.

(1.1) "International Telegraph and Telephone Consultative Committee's" (CCITT)
(1.2) "Telecommunications Industry Association" (TIA)

1.3.3.2 Medidas objetivas da qualidade da fala

O sistema de audição humano é o principal avaliador da qualidade e do desempenho de um codificador na preservação da inteligibilidade e da naturalidade. Enquanto extensivos testes subjetivos de audição provêem a mais precisa avaliação dos codificadores da fala, eles podem gastar muito tempo e serem incompatíveis com os propósitos práticos. Medidas objetivas podem fornecer uma imediata estimação confiável da qualidade perceptiva de um algoritmo de codificação [9].

Medidas Objetivas da Distorção no Domínio do Tempo

As medidas objetivas mais utilizadas no domínio do tempo são a relação sinal ruído "*Signal-to-Noise Ratio*" (SNR) e a relação sinal ruído segmental "*Segmental SNR*" (SEGSNR). Elas são sensíveis as variações de ganho e dos atrasos entre o sinal original e o sinal reconstruído. Elas também não avaliam completamente as propriedades do sistema de audição.

Relação sinal ruído ("*Signal-to-Noise Ratio*" (SNR)): é a medida objetiva mais comum para a avaliação do desempenho de um algoritmo de codificação. A SNR é definida como a razão da energia média da fala pela energia média do erro do sinal,

$$SNR = \sum_{n=-\infty}^{\infty} s^2(n) \Big/ \sum_{n=-\infty}^{\infty} \left(s(n) - \hat{s}(n)\right)^2 \tag{1.1}$$

usualmente ela é expressa em decibéis por:

$$SNR_{dB} = 10 \log_{10} SNR \tag{1.2}$$

onde $s(n)$ é o sinal da fala de entrada (ou sinal original) e $\hat{s}(n)$ é o sinal de saída (ou sinal reconstruído). A SNR é uma medida de longo termo para a precisão da reconstrução e como tal tende a camuflar os ruídos de reconstrução particularmente para os sinais de baixo nível.

Relação sinal ruído segmental ("*Segmental SNR*" (SEGSNR)): As variações temporais do desempenho podem ser detectadas e melhor avaliadas usando uma relação sinal ruído em curto termo (em uma base de quadro por quadro). A medida baseada por quadro é denominada de relação sinal ruído segmental. A SEGSNR é obtida em decibéis por

$$SEGSNR_{dB} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\frac{\sum_{n=m_j-N+1}^{m_j} s^2(n)}{\sum_{n=m_j-N+1}^{m_j}\left(s(n)-\hat{s}(n)\right)^2}\right] \tag{1.3}$$

onde M é o número de quadros, N é o número de amostras em um quadro e $m_0, m_1, \cdots, m_{M-1}$ são os índices de tempo para a última amostra de cada quadro. Para cada quadro (tipicamente de 15 a 25 ms), uma medida SNR é calculada e a medida final é obtida pela média dessas medições sobre todos os quadros da forma de onda. A vantagem de usar $SEGSNR_{dB}$ sobre a SNR convencional é que calculando a média sobre os valores de SNR_{dB} no domínio logaritmo ocorre uma melhor ponderação para os segmentos surdos da fala com baixa energia e assim a $SEGSNR_{dB}$ correlaciona melhor com a qualidade perceptiva do que a SNR.

Medidas Objetivas da Distorção no Domínio da Freqüência

No domínio da freqüência, o espectro de codificação "*Linear Predictive Coding*" (LPC) do sinal original é comparado com o espectro LPC do sinal quantizado ou interpolado. Uma medida da distorção espectral deveria ser capaz de mensurar as discrepâncias ou a diferença entre dois espectros que influenciam na percepção do som. Nas situações que seguem, as disparidades entre os envelopes espectrais original e o codificado podem levar a sons que perceptivamente são foneticamente diferentes:
- Os formantes dos envelopes espectrais, original e codificado, ocorrem em freqüências significantemente diferentes;
- A largura de faixa dos formantes nos envelopes espectrais difere significantemente.

A seguir é apresentada uma breve descrição dos tipos de medidas de distorções espectrais:

- *Medida de Itakura*: A medida de Itakura [10] geralmente corresponde melhor à qualidade perceptiva da fala. Também conhecida como "*likelihood ratio distance measure*", ela é a medida mais usada nos vocoders "*Linear Predictive*" (LP). Esta medida é baseada na similaridade entre os modelos somente - pólos das formas de ondas de referência e codificada. A medida de distância é calculada entre os conjuntos dos parâmetros LP estimados sobre quadros sincronizados (tipicamente a cada 15-30 ms) na fala original e codificada. A medida de Itakura é fortemente influenciada pelo erro espectral devido à falta de emparelhamento da

localização dos formantes, enquanto os erros no emparelhamento dos vales espectrais não contribuem de forma intensiva na distância. Isto é desejável, uma vez que o sistema auditivo humano é mais sensível a erros na localização e na largura de faixa do formante do que nos vales espectrais entre os picos.

- *Medida de Distorção Espectral Logarítmica* ("*Log Spectral Distortion Measure*"): A medida de distorção usada com maior freqüência é também denominada de "*Root Mean Square log spectral distortion*, (RMS), ou simplesmente distância espectral ("*spectral distance*"). A distância espectral para um dado quadro é definida como a diferença rms entre o "*LPC log power spectrum*" original e o "*LPC log power spectrum*" quantizado ou interpolado. Usualmente é calculada a média da distorção espectral sobre uma grande quantidade de quadros, e ela é utilizada como medida do desempenho da quantização ou da interpolação.

- *Medida da Distância Euclidiana Ponderada* ("Weighted Euclidean Distance Measure"): Esta medida é executada no domínio "*Line Spectral Frequency*" (LSF) ou no domínio do módulo espectral.

Outras medidas freqüentemente mencionadas na literatura incluem "*log-area ratio measure*", distância cepstral "*cepstral distance*" e "*articulation index*".

1.3.3.3 Medidas objetivas de predição da qualidade subjetiva da fala

Medidas objetivas em geral são sensíveis às variações dos ganhos e dos atrasos. Mais importante, elas tipicamente não levam em conta as propriedades espectrais do ouvido [11]. Por outro lado, avaliações subjetivas formais, podem ser longas e dispendiosas. Recentes esforços na avaliação da qualidade da fala têm estudado sobre o desenvolvimento de procedimentos e avaliação de testes automáticos e de medidas objetivas que sejam capazes de predizer a qualidade subjetiva da fala [3, 4, 5, 6]. Eles usam dois sinais como entrada, o sinal original (como padrão de referência) e o sinal de saída após sua transição pelo codificador da fala em teste como mostrado na Figura 1.2.

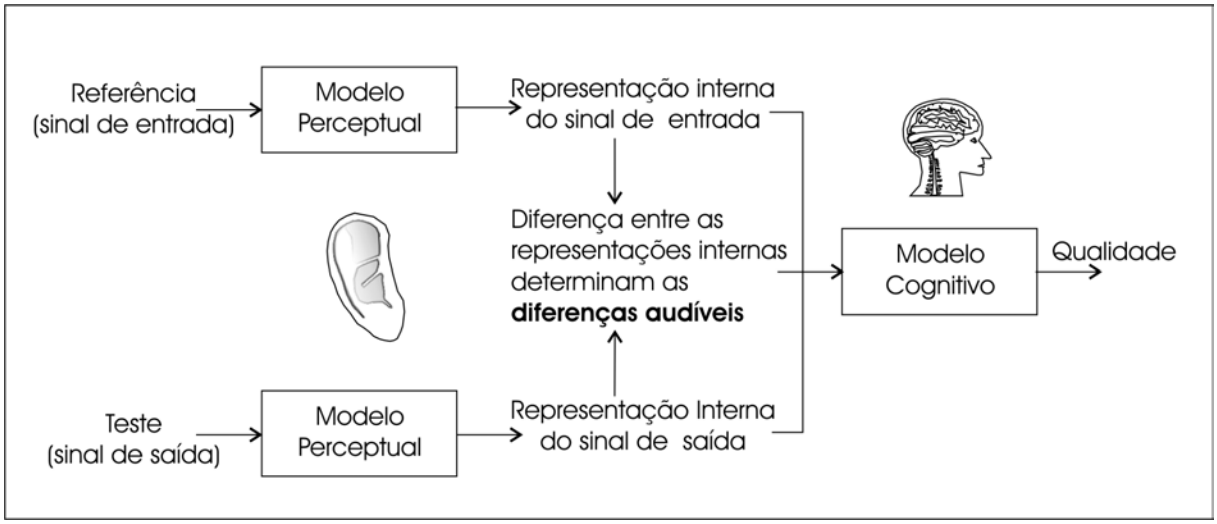

Figura 1.2 – Representação esquemática do modelo básico das medidas objetivas da qualidade da fala baseado nas medidas da qualidade perceptiva.

O processamento do sinal dentro das medidas perceptuais pode ser estruturado em três etapas: pré-processamento [12 e 13], modelamento psico-acústico, e modelamento cognitivo. O modelamento cognitivo é aquele que mais se diferencia dos métodos objetivos. O modelo cognitivo avalia o julgamento subjetivo da qualidade da fala. A avaliação é realizada pela determinação da distância perceptiva entre o sinal medido e a referência e então por criar uma figura de mérito que descreve a qualidade da fala. A figura de mérito é geralmente uma função não linear do valor MOS determinado subjetivamente. Com o intuito de se conseguir um estimador para o valor MOS, é necessário mapear o resultado objetivo para a escala MOS.

Os algoritmos mais conhecidos para a avaliação objetiva da qualidade da fala baseada no modelo de percepção psico-acústico dos sons são: "*Bark Spectral Distance*" (BSD) [9], "*Perceptual Speech Quality Measure*" (PSQM) recomendação P.861 da ITU-T[(1.3)] [6] e "*Perceptual Evaluation of Speech Quality*" (PESQ) recomendação P.862 da ITU-T [5].

Neste trabalho são utilizadas a SEGSNR como método de avaliação objetiva e o PESQ como método de avaliação objetiva da qualidade perceptiva. Estes métodos são aplicados aos sinais da fala na avaliação dos sinais da fala sintetizados e aos sinais residuais da fala na avaliação dos sinais residuais reconstruídos.

1.4 Objetivos e propósitos desta tese

A pesquisa está relacionada com a codificação da fala utilizando a técnica de codificação por interpolação de ondas, também denominada de "Waveform Interpolation Speech Coding"

(1.3) "*Telecommunication Standardization Sector of International Telecommunication Union*" (ITU-T)

(WI), técnica WI. Esta técnica ultimamente tem sido muito pesquisada e promete avanços na codificação da fala em taxas baixas de bits. O algoritmo desenvolvido neste trabalho foi baseado principalmente nos trabalhos de E. Choy [14] e também auxiliado pelos trabalhos de W. B. Kleijn [15]. O trabalho de Choy [14] descreve um algoritmo de um codificador WI, em um nível de maior abstração, em diagrama de blocos, operando na faixa telefônica (faixa estreita) para atingir uma qualidade da fala próxima a "*toll-quality*" em 4 kbits/s.

Na literatura científica não existem algoritmos disponíveis para a codificação da fala utilizando a técnica WI. Alguns centros de pesquisa no mundo patrocinados por grandes empresas de telecomunicações estudaram e desenvolveram seus próprios algoritmos para a codificação da fala usando a interpolação de ondas, os quais são mantidos em segredo comercial. Devido a isso, nesta tese o propósito é iniciar à pesquisa com a técnica WI tornando disponível os resultados obtidos aos pesquisadores da área acadêmica.

Um sistema de codificação da fala usando a técnica WI é composto por duas camadas, a *camada externa* – onde é realizada a análise e a síntese do sinal da fala e a *camada interna* – onde é realizada a codificação (ou quantização) e a decodificação dos parâmetros. Este trabalho estuda a camada externa, com a intenção de analisar e sintetizar o sinal da fala usando a técnica WI, e portanto sem codificar (ou quantizar) os parâmetros. Assim, o objetivo da tese é descrever, visualizar e implementar um algoritmo para um analisador/sintetizador WI. O trabalho portanto está focalizado no *estudo e na descrição da técnica de codificação de sinais da fala utilizando interpolação de ondas, técnica WI, e na visualização e na implementação de um algoritmo para simular um sistema convencional básico de codificação WI, analisador/sintetizador WI.*

Assim foi implementado o algoritmo de um codificador/decodificador WI básico e convencional. *Básico* por operar na camada externa, o *analisador/sintetizador WI,* onde os parâmetros são transmitidos sem compressão, e *convencional* por não procurar alcançar a reconstrução perfeita. Portanto, o interesse é executar a análise do sinal da fala obtendo os parâmetros, transmití-los sem codificação e fazer a reconstrução (ou síntese) do sinal da fala. Este sistema será denominado de *sistema de análise – síntese WI (padrão)*.

Este trabalho também inicia a pesquisa sobre as inovações nos processos de localização e extração das formas de ondas características, as CW's, que são obtidas por interpolação nas posições regulares durante o estágio de análise (na codificação), com o objetivo de se obter uma melhorar reconstrução do sinal durante o estágio de síntese. O sistema resultante a partir desta pesquisa será denominado de *sistema de análise – síntese WI (com interpolação)*.

Com este trabalho a técnica WI fica mais acessível e também inicia e possibilita a continuidade da pesquisa para completar e/ou aperfeiçoar o sistema WI de codificação da fala a nível acadêmico neste grupo de pesquisa.

1.5 Organização da Tese

O trabalho foi desenvolvido em 4 fases:
1ª fase - *Estudo introdutório sobre a codificação do sinal da fala compreendendo:*
- a motivação para a codificação de sinais da fala;
- uma introdução sobre a produção, percepção e medidas da qualidade da fala;
- uma visão geral sobre os codificadores, as características principais de cada técnica, e os atributos de codificadores da fala;
- e a descrição das técnicas e os métodos auxiliares de processamento para a análise e a síntese de sinais da fala utilizados no trabalho.

2ª fase - *Estudo e descrição da técnica de codificação de sinais da fala utilizando interpolação de ondas, técnica WI.*
3ª fase – *Visualização* e *Implementação de um alg*oritmo para simular um sistema convencional básico de codificação WI, analisador/sintetizador, ou *sistema de análise – síntese WI (padrão)*.
4ª fase – *Resultados da simulação e avaliação dos sistemas de análise – síntese WI (padrão)* e da versão preliminar do *sistema de análise – síntese WI (com interpolação) com alguns sinais da fala.*

1.5.1 Os capítulos da tese

Na Tese, as fases estão relacionadas em 6 capítulos e 1 apêndice que podem ser resumidos como seguem:
- no capítulo 1 é apresentada uma introdução à codificação da fala destacando sua importância nos sistemas de telecomunicações, a produção e a percepção da fala, a performance dos codificadores, os objetivos e o escopo desta pesquisa, e a organização desta tese;
- no capítulo 2 são descritas as técnicas de processamento de sinais utilizadas neste trabalho, a análise de predição linear (LPC) e alguns procedimentos de processamento digitais de sinais da fala que auxiliam na preparação do sinal da fala para os processos de codificação e decodificação utilizados nesta tese.
- no capítulo 3 são apresentadas de forma resumida as classificações e a visão geral sobre as principais técnicas de codificação da fala com a indicação de alguns codificadores padronizados para os sistemas de telecomunicações;
- o capítulo 4 apresenta a descrição da técnica de codificação da fala por interpolação de ondas com a apresentação em detalhes de alguns procedimentos e algoritmos que facilitam a visualização de um algoritmo geral para a implementação de um sistema de codificação WI, convencional e básico (analisador/sintetizador), ou *sistema de análise – síntese WI (padrão)*;

- o capítulo 5 apresenta os *resultados e* a *avaliação da simulação* com o sistema WI implementado, o sistema de *análise – síntese WI (padrão)* relacionando: os sinais da fala, original e sintetizado (ou reconstruído); e os sinais residuais da fala, original (no lado da análise) e o reconstruído (no lado da síntese). Também apresenta os propósitos, os resultados e a avaliação para a versão preliminar do *sistema de análise – síntese WI (com interpolação)*.
- no capítulo 6 são apresentados as conclusões, as contribuições da tese e os trabalhos futuros.
- o apêndice A apresenta propostas para melhoria na localização e na interpolação das formas de ondas características, as CW's, em posições regulares. Parte desta proposta foi aplicada ao sistema de análise – síntese WI (padrão) tendo resultado, após aprimoramentos, na versão preliminar do sistema de análise – síntese WI (com interpolação).

1.6 Considerações Finais deste Capítulo

Neste capítulo foi apresentada uma introdução à codificação da fala destacando sua importância nos sistemas de telecomunicações, a produção e a percepção da fala, as medidas para verificação da performance dos codificadores, os objetivos e o escopo desta pesquisa e a organização desta tese.

Capítulo 2

PROCESSAMENTO DO SINAL DA FALA

2.1 Introdução

Este capítulo trata do processamento dos sinais da fala e das técnicas e dos métodos auxiliares empregados neste trabalho.

Na seção 2.2 é fornecida uma visão geral sobre o processamento digital do sinal da fala, mostrando os objetivos do processamento digital, a representação dos sinais digitais e as aplicações gerais na comunicação dos sinais da fala.

Nas seções 2.3 e 2.4 são apresentados uma visão dos processos fundamentais de produção da fala e o modelo matemático discreto no tempo que é utilizado para representar os sinais amostrados da fala. O modelo discreto no tempo para a produção da fala servirá de base para a aplicação das técnicas de processamento digital.

Na seção 2.5 são apresentados os métodos e as técnicas de processamento de sinais utilizados neste trabalho, que têm como objetivo principal a análise (extração dos parâmetros), preparando o sinal para a codificação, e a síntese do sinal da fala (a partir dos parâmetros extraídos). A codificação da fala por predição linear "*Linear Predictive Speech Coding*" (LPC) é uma técnica que envolve: a *análise LP* do sinal da fala, *"Linear Prediction Analysis" (LP)*, como o método de extração dos vetores LPC (ou coeficientes LPC) e como método de cálculo do sinal residual; e também a *síntese LP "Linear Prediction Synthesis" (LP)* como método de reconstrução do sinal da fala. Depois é apresentado um resumo dos procedimentos para a aplicação da análise LP e da síntese LP. Em seguida são apresentados o processo de conversão dos coeficientes LPC para coeficientes "*Line Spectral Frequency*" (LSF) que são mais adequados para os processos de interpolação e de quantização e, o processo de conversão dos coeficientes LSF para coeficientes LPC. A expansão da largura de faixa e a pré-ênfase auxiliam na preparação do sinal da fala para uma análise mais eficiente dos coeficientes de predição linear, e a de-ênfase reverte o processo de pré-ênfase, após a síntese LP do sinal da fala.

2.2 Processamento de Sinais

O problema geral da manipulação e do processamento da informação é representado na Figura 2.1. No caso dos sinais da fala o locutor humano é a fonte de informação. As medidas ou as

observações são obtidas geralmente da forma de onda acústica. O processamento envolve primeiramente a obtenção da representação do sinal baseada em um dado modelo e então a aplicação de alguma transformação em algum nível mais elevado, colocando o sinal em uma forma mais conveniente para a sua manipulação. O último passo é a extração e a utilização das informações, que podem ser realizadas ou pelos ouvintes humanos ou pelas máquinas.

O processamento dos sinais da fala envolve duas tarefas:

1ª) É o meio para se obter uma representação geral dos sinais da fala, em forma de onda acústica, ou em forma paramétrica;

2ª) O processamento dos sinais atua com a função de auxiliar no processo de transformação da representação do sinal em formas alternativas, que são menos genéricas, porém mais apropriadas para aplicações específicas.

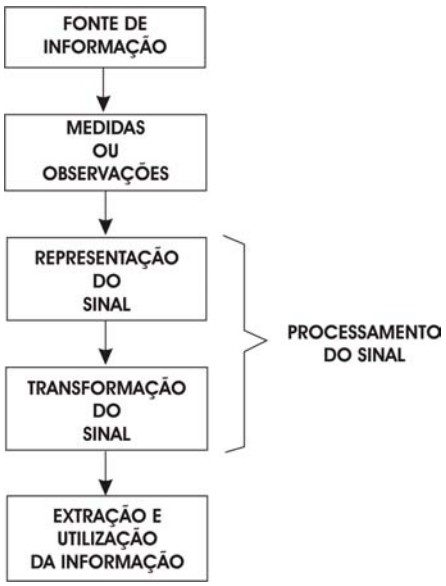

Figura 2.1 – Visão geral do processamento e manipulação da informação [16].

2.2.1 Processamento digital de sinais

O processamento digital de sinais se interessa pela obtenção da representação dos sinais digitais, e pela teoria, projetos e implementações dos procedimentos numéricos para o processamento da representação discreta.

"Os objetivos do processamento digital de sinais são idênticos ao processamento de sinais analógicos. A principal razão da utilização das técnicas de processamento digitais, no contexto especial da comunicação da fala e provavelmente a mais importante, é o fato de que as funções do processamento de sinais extremamente sofisticadas possam ser utilizadas

usando as técnicas digitais. Outra razão é que com uma adequada codificação, a fala na forma digital pode ser transmitida de forma confiante sobre muitos canais ruidosos. Além disto, o sinal da fala na forma digital tem forma idêntica a de outros tipos de dados. Assim, na transmissão na rede não existe a necessidade da distinção da fala e de outros dados exceto na decodificação. Com respeito à transmissão dos sinais de voz requerendo segurança, a representação digital tem uma vantagem distinta sobre os sistemas analógicos. Para a segurança, os bits de informação podem ser embaralhados na transmissão, e de alguma forma podem ser desembaralhados na recepção" [16]. Também os avanços na microeletrônica favorecem as implementações digitais, permitindo a implementação de algoritmos com maior complexidade em aparelhos cada vez menores. Portanto existem muitas razões para concluir que, salvo algumas exceções, a representação discreta do sinal é muito mais interessante do que na forma contínua.

A aplicação das técnicas de processamento de sinais digitais em problemas de comunicação da fala pode ser vista sob três tópicos principais [16]: (1°) a representação dos sinais da fala na forma digital; (2°) a implementação das técnicas de processamento sofisticadas; e (3°) - as áreas de aplicações (voz na Internet - "Voice Over Internet Protocol" (VoIP), sistemas de telefonia fixa e móvel, e também como mostra a Figura 2.3).

"O interesse fundamental é a representação dos sinais da fala na forma digital. Esta representação é feita utilizando-se o *teorema da amostragem*, que estabelece que um sinal filtrado, limitado em faixa, pode ser representado por amostras obtidas periodicamente no tempo, desde que a amostragem seja realizada em uma taxa adequada. Assim o processo de amostragem é realçado em toda a teoria e nas aplicações de processamento digital de sinais" [16].

"A Figura 2.2 mostra que as *representações discretas* podem ser classificadas em dois grupos gerais. No primeiro, denominado de *representações da forma de onda*, a forma de onda do sinal analógico é preservada no processo de amostragem e de quantização. No segundo, denominado de *representações paramétricas*, o interesse é a representação do sinal da fala como a saída de um modelo de produção da fala, síntese. O primeiro passo na obtenção da representação paramétrica é muitas vezes a representação digital da forma de onda na qual o sinal é amostrado e quantizado e então processado para a obtenção dos parâmetros do modelo de produção da fala. Os parâmetros deste modelo são convenientemente classificados ou como *parâmetros de excitação*, isto é, relacionados à fonte de produção dos sons da fala, e *parâmetros do trato vocal*, relacionados aos sons individuais da fala" [16].

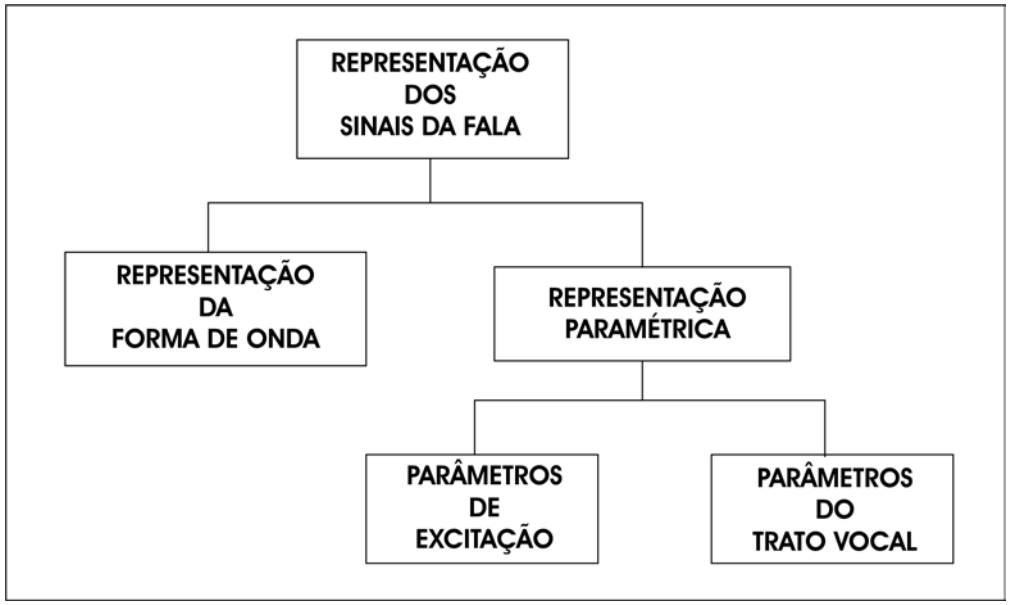

Figura 2.2 – Representações dos sinais da fala [16].

A Figura 2.3 mostra algumas das áreas de aplicações na comunicação da fala [16].

Figura 2.3 – Algumas aplicações típicas da fala na comunicação [16].

2.3 O Processo de Produção da Fala

O processo de produção da fala envolve os centros da fala no cérebro, o centro respiratório na base do cérebro, o sistema respiratório, a cavidade torácica, as estruturas da laringe, a faringe, as cavidades nasais, o nariz, as estruturas e partes da boca e os músculos faciais. Os mecanismos de produção da fala (ou vocalização) envolvem a respiração, a fonação, a ressonância e as articulações.

O processo da fala começa com a expiração do ar produzido pelo mecanismo respiratório da expansão e da contração dos pulmões. A energia dirigida para a produção da fala, gerada pelo mecanismo da expiração varia com a fisiologia individual, hábitos de respiração e treinamento para correção onde for necessário.

No processo de fonação o fluxo de ar é vibrado na laringe pelas cordas vocais, que são pregueadas ao longo das paredes laterais da laringe e são esticadas lateralmente entre dois conjuntos de cartilagens. A pressão do ar primeiro empurra as cordas vocais separando-as. Isto cria um vácuo parcial entre as cordas vocais que podem ser puxadas e juntadas novamente e assim, continuam com um determinado padrão de vibração.

A coordenação, tamanho, elasticidade, saúde das complexas estruturas da laringe, as cordas vocais e os músculos intrínsecos e extrínsecos da laringe atuam sobre o fluxo de ar e em combinação com a forma e o posicionamento destas cordas vocais não apenas produzem sons, mas também causam a individualização de cada som. O imediato resultado da vibração das cordas vocais é o tom fundamental, ou *pitch* da voz. Em termos físicos a freqüência de vibração (se ela for alta a voz é aguda, e se for baixa a voz é grave, ou intermediária) como atributo primário das cordas vocais corresponde ao número de sopros por segundo. Esta freqüência, resultando no *pitch* da voz, é determinada por fatores estáveis e fatores variáveis. Fatores estáveis estão relacionados com as dimensões da laringe, sexo, idade, e tipo de corpo. Fatores variáveis incluem a tensão das cordas vocais, a força de fechamento da glote e a pressão do fluxo de ar expirado. O *pitch* da voz aumenta ou diminui com o aumento ou a diminuição da tensão externa dos músculos da laringe. Então, o tom de voz reconhecível fundamentalmente é determinado primariamente pela freqüência de vibração das cordas vocais, é uma função da laringe. Estas funções são determinadas por características congênitas e adquiridas nos comportamentos por aprendizagem com o uso da voz.

O processo de ressonância, que é a amplificação do som fundamentalmente no modo de falar (forma de expressar) envolve a faringe, a boca, o nariz, as cavidades nasais, e a cavidade torácica. A qualidade da ressonância que pode ir da estridência a inaudibilidade, também depende de fatores estáveis e de fatores variáveis, que consideram as condições físicas e os comportamentos aprendidos e está relacionado com o intento individual e sua personalidade, bem como seus comportamentos da fala. Isto também está relacionado com a força de expiração do ar e as dimensões da cavidade torácica. Devido aos vários efeitos e ao uso destas partes do mecanismo de ressonância, certos tipos de fala são exibidos, tais como a nasalidade, que representa um tipo de som com o acoplamento da cavidade nasal ao trato vocal para a ressonância ou uma boa projeção do som, que usa a cavidade torácica tão bem quanto outros órgãos.

O processo de articulação constitui a formação de sons amplificados em palavras, pelo movimento dos lábios, língua, do palato mole da boca e relacionados aos músculos faciais. Mais ainda, a qualidade de uma dada linguagem pode requerer diferentes formas de articulação porque a zona linguodental (entre a ponta da língua e os dentes) pode ser usada diferentemente em uma ou outra linguagem. Universalmente, entretanto, os sons da fala devido a articulação são agrupados como sons orais ou sons nasais, e ambos relacionam as

condições estruturais e comportamentos da fala tão bem como relacionam as disciplinas de lingüística, desenvolvimento da fala e pronunciação.

2.3.1 Sistema fonador

A Figura 2.4 mostra que o sistema fonador pode ser dividido em três subsistemas:

1. **Subsistema respiratório**: A respiração atua como fonte de energia mecânica para iniciar o processo da fonação.

2. **Subsistema laringeal**: Responsável pela produção da energia sonora como fonte de excitação (transforma a energia mecânica do fluxo de ar em ondas acústicas).

3. **Subsistema supra-laringeal**: responsável pela modificação, modulação do som produzido pela laringe (pela ressonância e pelas articulações do trato vocal).

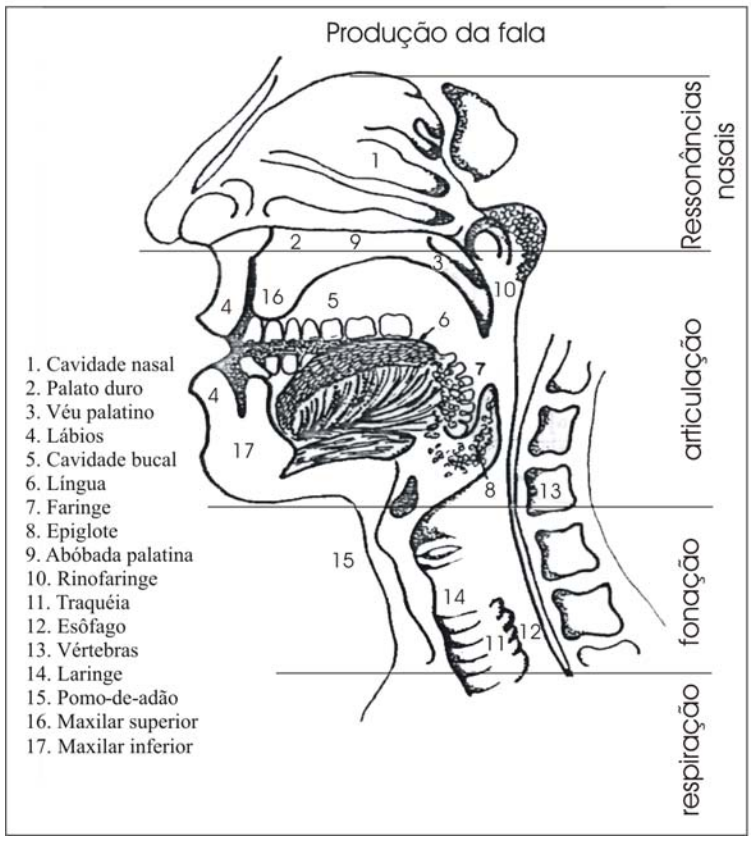

Figura 2.4 – Aparelho fonador.

Subsistema respiratório: é composto pelos *pulmões*, *brônquios* e *traquéia* que são órgãos responsáveis pela produção do fluxo de ar (pulmões) e pela condução do mesmo (brônquios e traquéia, canais condutores).

Subsistema laríngeo: constituído pela *laringe*, que situada na parte superior da traquéia, é o mais importante órgão da fonação. Nela localizam-se a *glote*, a *epiglote* (válvula elástica que tapa a glote durante a deglutição) e as *cordas vocais*. A glote é uma pequena abertura em forma triangular, situada entre duas pregas musculares elásticas, conhecidas como cordas vocais. Estas são pregueadas ao longo das paredes laterais da laringe e esticadas lateralmente entre dois conjuntos de cartilagens. O fluxo de ar vindo dos pulmões pode encontrar a glote aberta ou fechada, ou seja, as cordas vocais com os bordos afastados ou juntos. Se a glote se abrir, ou estiver aberta, o ar passa livremente, sem vibrar as cordas vocais: nesse caso, o *fonema* produzido é *surdo* (ex.: */p/*). Caso, contrário, a glote fechada, o fluxo de ar encontra resistência a sua passagem, e as cordas vocais, que estão com as bordas aproximadas (juntas), vibram e o *fonema* produzido é então *sonoro* (ex.: */a/*).

Subsistema supra-laringeal: é constituído pela *faringe* (conexão do esôfago e da boca), *boca* (cavidade oral que se estende da úvula para os lábios) e *fossas nasais*. É limitado também pelas seguintes estruturas: epiglote, maxilar inferior, língua, úvula, palato mole, palato duro, dentes e lábios.

A úvula que é um apêndice flexível do palato mole (ou véu palatino) tem a função de controlar a passagem do fluxo de ar. Levantando a úvula contra a parede posterior da faringe, intercepta a passagem do ar para as fossas nasais e o fluxo escoa-se pela cavidade bucal e o *fonema* é *oral* (ex.: */o/*---b**o**la). Abaixando a úvula, permite que parte do fluxo escape pelas fossas nasais, produzindo então um *fonema nasal* (ex.: */õ/*---p**õ**e).

Estes conjuntos de órgãos, (cavidades supra-laríngeanas) atuam como caixas de ressonância para fonemas orais e nasais, uma vez que a cavidade bucal consegue variar bruscamente sua forma e volume, devido ao movimento dos órgãos como a língua e os maxilares, podendo ainda ser acoplada às cavidades nasais. A onda acústica, fluxo de ar vindo da laringe, é então modificada (modulada) para moldar os fonemas.

A cavidade bucal é então a grande responsável pela diversificação do som, pelos movimentos da língua e dos lábios, e da maior ou da menor separação entre as bochechas.

2.3.2 O Sinal da fala

A fala é essencialmente um instrumento de comunicação [16]. Existem várias maneiras de caracterizar o potencial de comunicação da fala. Uma forma altamente quantitativa é em

termos da teoria da informação, introduzida por Shannon. De acordo com esta teoria, a fala pode ser representada em termos de seu conteúdo de mensagem ou da informação. Um modo alternativo de caracterizar a fala é pelo sinal transportando a informação da mensagem, ou seja, pela forma de onda temporal do sinal acústico.

Ao considerar o processo de comunicação da fala, é útil começar pelo pensamento de uma mensagem representada em uma forma abstrata no cérebro de um locutor. A fala é produzida por um processo complexo, a mensagem é convertida ultimamente para um sinal acústico. A informação na mensagem é primeiramente convertida em um conjunto de sinais neurais que controlam o mecanismo articulatório, ou seja, o movimento da língua, os lábios, as cordas vocais, etc. Os articuladores em resposta aos sinais neurais executam uma seqüência de movimentos, que resultam em uma forma de onda acústica que contém a informação original da mensagem [16].

Em sistemas de comunicação, o sinal da fala é transmitido, armazenado e processado de várias maneiras. Os interesses técnicos levam a uma ampla variedade de representações do sinal da fala. Em geral, qualquer uma delas mantém a preocupação em preservar o conteúdo da mensagem, representando o sinal da fala em uma forma conveniente ou em uma forma flexível para que as modificações, ou processamento, possam ser feitos no sinal da fala sem degradar o conteúdo da mensagem.

A representação do sinal da fala deve ser de tal forma que o conteúdo da mensagem seja facilmente extraído pelos ouvintes humanos ou automaticamente pelas máquinas.

2.4 O Modelo Discreto no Tempo para a Produção da Fala

Para aplicar as técnicas de processamento de sinais em problemas da comunicação da fala é essencial o entendimento dos processos fundamentais da produção da fala. Esta seção mostra que existe uma variedade de formas de representar o sinal da fala. Especificamente será mostrado como obter os modelos discretos no tempo para a representação dos sinais da fala amostrados. Estes modelos servirão de base para a aplicação das técnicas de processamento digital. Várias referências excelentes existem para detalhar mais o assunto, como é o caso das obras de Fant e Flanagan [16].

Os sinais da fala são compostos por uma seqüência de sons, que em conjunto com as transições entre eles servem como a representação simbólica da informação [16]. O arranjo desses sons é governado pelas regras da linguagem, sendo que o estudo dessas regras e de suas implicações no processo da comunicação humana, é tratado pela **lingüística**. O estudo e a classificação das unidades de sons da fala são desempenhados pela **fonética** [16].

O objetivo nesta seção é a apresentação de um modelo matemático, desenvolvido por Gunnar Fant [16], que representa, mesmo que superficialmente, os mecanismos de produção

da fala e de propagação do som, que serve como base para a análise e a síntese deste mesmo sinal.

2.4.1 O mecanismo de produção da fala

Nas seções anteriores deste capítulo foram apresentados os órgãos, as características e funcionamento do sistema fonador.

No estudo do processo de produção da fala, é útil abstrair as características importantes do sistema físico de uma forma que leve ao mais realístico e tratável modelo matemático. A Figura 2.5 apresenta um diagrama esquemático do **sistema vocal** (análogo ao aparelho vocal, mostrado na Figura 2.4) onde se destacam três subsistemas independentes, porém interligados:

1. **Sistema sub-glotal** - composto pelos pulmões, brônquios e pela traquéia, serve como a fonte de energia para a produção da fala;

2. **Trato vocal** - começa na abertura das cordas vocais (glote), estendendo-se até aos lábios. É onde o sinal acústico é gerado (cordas vocais), e encontra os obstáculos necessários ao seu modelamento. Para os homens, atinge em média, cerca de 17 cm de comprimento, com uma área da seção transversal que pode variar de zero (completo fechamento) a 20 cm², em média;

3. **Trato nasal** - começa no véu palatino (úvula) e estende-se até as narinas. É um dos responsáveis pela emissão dos sons nasais da fala, quando o véu palatino é abaixado e o trato nasal é acoplado ao trato vocal.

A fala é simplesmente a onda acústica resultante que é radiada deste sistema quando a corrente de ar é expelida dos pulmões e resultam em um fluxo de ar que é perturbado por alguma constrição em algum lugar no trato vocal.

A operação do sistema vocal é dividida em duas funções, excitação e modulação, como mostrado na Figura 2.6.

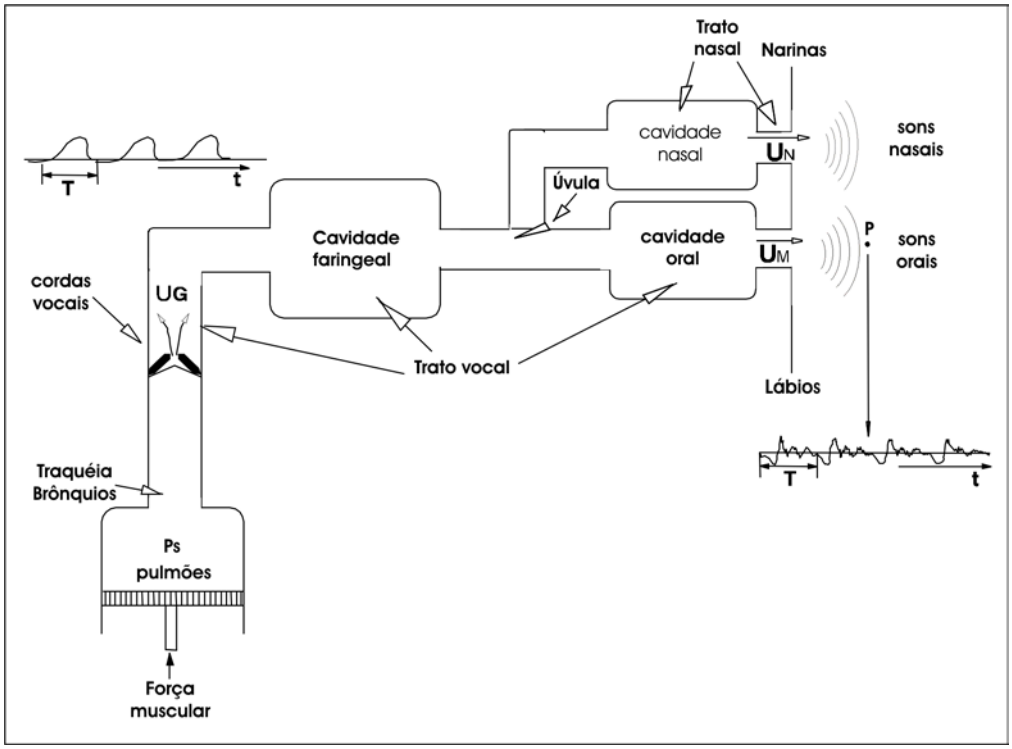

Figura 2.5 – Diagrama esquemático do aparelho fonador – produção da fala humana adaptada de [17].

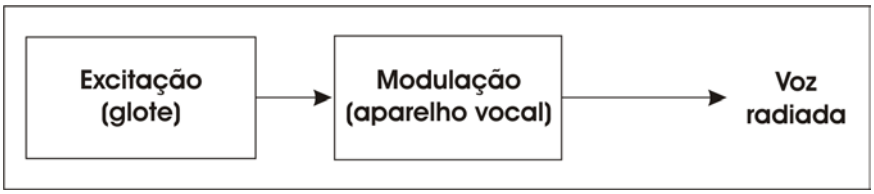

Figura 2.6 – Modelo de produção da fala (Excitação e Modulação) [16].

I. Excitação - É comum classificar os sons da fala de acordo com seu modo de excitação. Existem dois tipos elementares de excitação do trato vocal [17]: **sons sonoros e sons surdos.** Três outros tipos de excitação, que são realmente as combinações de sons sonoros, surdos e silêncio, são delineados com o propósito de classificação e modelamento. Eles são: sons mistos, sons oclusivos, e silêncio.

(a) **Os sons sonoros**: são produzidos como resultado da excitação por uma série de pulsos de ar quase periódicos gerados pelas cordas vocais. A freqüência fundamental das vibrações das cordas vocais, também chamada de ***pitch***, é determinada pela massa e tensão das cordas vocais e pela pressão do ar sub-glotal. Ambas, a tensão nas cordas vocais e a pressão do ar sub-glotal, variam durante o processo de produção da fala. A faixa de freqüência fundamental (*pitch* da voz) na fala dos adultos varia de 50 a 500 Hz [17].

(b) **Os sons surdos**: ocorrem quando a glote está aberta. Estando aberta, a corrente de ar passa livremente por ela sem que vibrem as cordas vocais e a fonte de excitação é movida (da laringe) para algum ponto de constrição ao longo da cavidade bucal. A constrição produz fluxos turbulentos de ar e a excitação, assim produzida, é um *ruído de espectro amplo* para excitar o aparelho vocal.

(c) **Os sons mistos**: um som pode ser simultaneamente sonoro e surdo. Por exemplo, o fonema /z/ na palavra 'zebra'.

(d) **Os sons oclusivos**: são produzidos pelo fechamento momentâneo do aparelho vocal impedindo a saída do ar. Com isto a pressão aumenta dentro do trato vocal e uma rápida liberação causa o movimento do ar como uma explosão (excitação transitória).

(e) **O silêncio**: é questionável se o silêncio é uma forma de excitação, mas é útil incluí-lo com o propósito de modelamento [17].

II. Modulação - A fala é modelada em função dos movimentos dos órgãos da voz na cavidade bucal. Os tratos vocais e nasais são mostrados na Figura 2.5 como tubos com área da seção transversal não uniforme. Os sons gerados propagam-se nesses tubos. *O espectro de freqüência é formado pelas freqüências seletivas do tubo*. Este efeito é similar aos efeitos de ressonância nos instrumentos de sopro. As *freqüências de ressonância no trato vocal são chamadas de* **freqüências formantes** ou simplesmente **formantes**. Elas dependem da forma e das dimensões do trato vocal; cada forma é caracterizada por um conjunto de freqüências formantes. A variação da forma do trato vocal resulta na formação de diferentes sons. Assim as propriedades espectrais do sinal da fala variam no tempo como varia a forma do trato vocal.

2.4.2 Modelo digital para sinais da fala

Um modelo matemático detalhado pode ser obtido a partir da acústica da produção da fala. A teoria acústica relaciona as características básicas do sinal da fala com a física da produção da mesma.

Os sons da fala são gerados por excitações de tipos diferentes e cada um resulta em um tipo distinto de saída. O aparelho vocal impõe suas ressonâncias sobre a excitação para produzir os diferentes sons da fala.

A idéia que surge é a representação da produção da fala em termos de um modelo terminal análogo como mostrado na Figura 2.7. Este modelo é um sistema linear que produz

uma saída com as propriedades desejadas quando controlado por um conjunto de parâmetros relacionados com o processo de produção da fala. O modelo é equivalente ao modelo físico em sua saída (término) produzindo formas de ondas análogas (o sinal da fala), mas em sua estrutura interna não imita a física da produção da fala. O interesse final é trabalhar com modelo terminal análogo discreto no tempo para a representação dos sinais amostrados da fala.

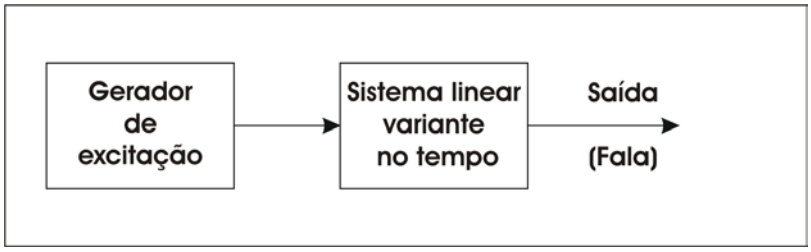

Figura 2.7 – Modelo esquematizado da produção da fala "modelo terminal análogo" [16].

Para produzir um sinal da fala, o modo de excitação e as propriedades de ressonância do sistema linear precisam variar com o tempo. A natureza desta variação foi vista na subseção 2.4.1 deste capítulo. As propriedades do sinal da fala muda lentamente com o tempo. Para muitos sons da fala é razoável supor que as propriedades gerais da excitação e do trato vocal permanecem fixas por períodos de 10 a 20 milissegundos. Desta forma, um modelo terminal análogo envolve um sistema linear variante no tempo excitado por um sinal cuja natureza básica muda de pulsos quase periódicos (em sinais sonoros) para ruídos aleatórios (em sinais surdos).

As características essenciais do modelo discreto no tempo dos tubos sem perdas são representadas na Figura 2.8.

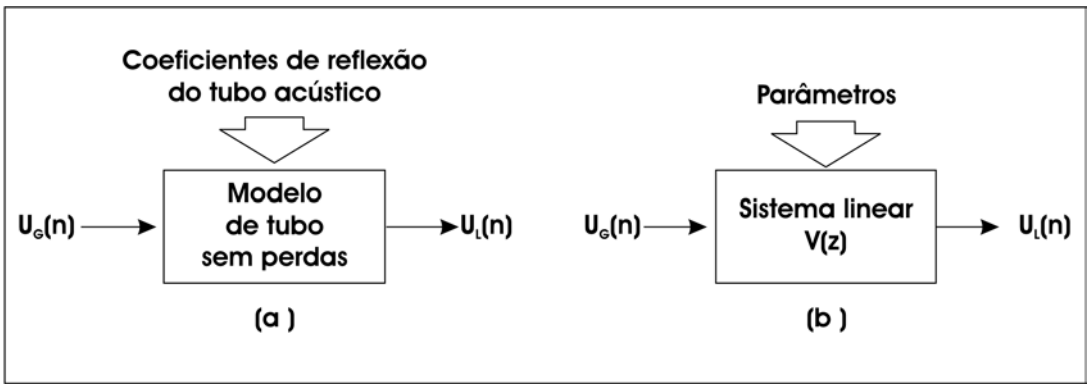

Figura 2.8 – (a) Diagrama de blocos representando o modelo dos tubos sem perdas; (b) Modelo terminal análogo [16]

No modelo da Figura 2.8, o trato vocal foi caracterizado por um conjunto de áreas, ou equivalentemente, pelos coeficientes de reflexão. A relação entre a entrada (U_G – vazão

volumétrica do ar através da glote) e a saída (U_L – vazão volumétrica do ar através dos lábios) pode ser representada por uma função de transferência, V(z) da forma

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k \; z^{-k}} \tag{2.1}$$

onde G (ganho total) e $\{\alpha_k\}$ (coeficientes de reflexão do tubo acústico) dependem da função área. Qualquer sistema tendo esta função de transferência produzirá a mesma saída, em resposta a uma determinada entrada. Assim, os modelos terminais análogos possuem a forma geral da Figura 2.8b.

Para um modelo ser completo deve incluir também a função de mudança de excitação e os efeitos da radiação do som nos lábios à resposta do trato vocal, $R(z)$.

No **trato vocal** as ressonâncias (formantes) da fala correspondem aos pólos da função V(z). Um modelo somente-polos é uma boa representação do efeito do aparelho vocal para a maioria dos sons; entretanto, a teoria acústica nos diz que os sons nasais e os fricativos requerem tanto ressonância quanto anti-ressonância (pólos e zeros). Nesses casos, pode-se incluir zeros na função de transferência.

Para o **efeito da radiação** nos lábios uma aproximação razoável é obtida com:

$$R(z) = R_o \; (1 - z^{-1}) \; . \tag{2.2}$$

A Figura 2.9 representa um modelo incluindo o trato vocal e os efeitos da radiação nos lábios.

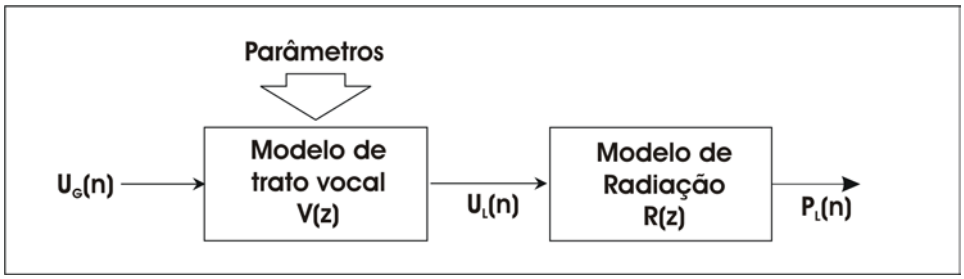

Figura 2.9 – Modelo terminal análogo incluindo os efeitos da radiação nos lábios [16].

Para completar o modelo terminal análogo, é necessário gerar um sinal apropriado de entrada (sinal de excitação) para o sistema do trato vocal. Relembrando que a maioria dos sons da fala pode ser classificada como sonoros ou surdos, foi visto neste capítulo em termos gerais que é requerida uma fonte que produza ondas em forma de pulsos quase periódicos ou em forma de ruído aleatório.

A Figura 2.10 esquematiza a geração do sinal de excitação para os sons sonoros, ou seja, a onda glotal. O gerador de trem de impulsos produz uma seqüência de impulsos unitários espaçados por um período fundamental (*período de pitch*) desejado. Este sinal excita um sistema linear cuja resposta impulsiva g(n) tem a forma de onda glotal desejada. O controle de ganho, A_v, controla a intensidade da excitação sonora.

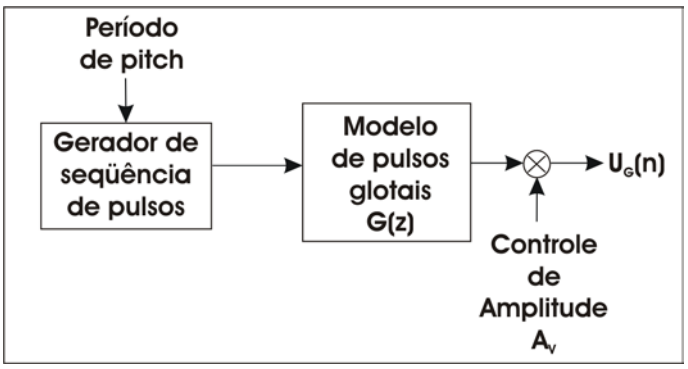

Figura 2.10 – Geração do sinal de excitação para os sons sonoros [16].

Rosenberg [16], em estudos do efeito da forma do pulso glotal na qualidade da fala, descobriu que a forma de onda glotal natural poderia ser substituída por uma forma de onda de pulso sintético como:

$$g(n) = \begin{cases} \frac{1}{2}\left[1 - \cos(\pi n / N_1)\right] & 0 \le n \le N_1 \\ \cos(\pi(n - N_1)/2N_2) & N_1 \le n \le N_1 + N_2 \\ 0 & \textit{outros casos} \end{cases} \qquad (2.3)$$

A transformada z da Equação (2.3), G(z), tem apenas zeros. Ela apresenta bons resultados com modelo de 2 pólos para G(z).

Para os sons surdos o modelo de excitação é muito mais simples. Tudo que é requerido é uma fonte de ruído aleatório e um parâmetro de ganho, A_N, para controlar a intensidade da excitação surda. Para os modelos discretos no tempo, um gerador de ruído aleatório produz uma fonte de ruído de espectro plano.

Um **modelo completo** com a junção dos elementos já descritos neste capítulo pode ser visto na Figura 2.11.

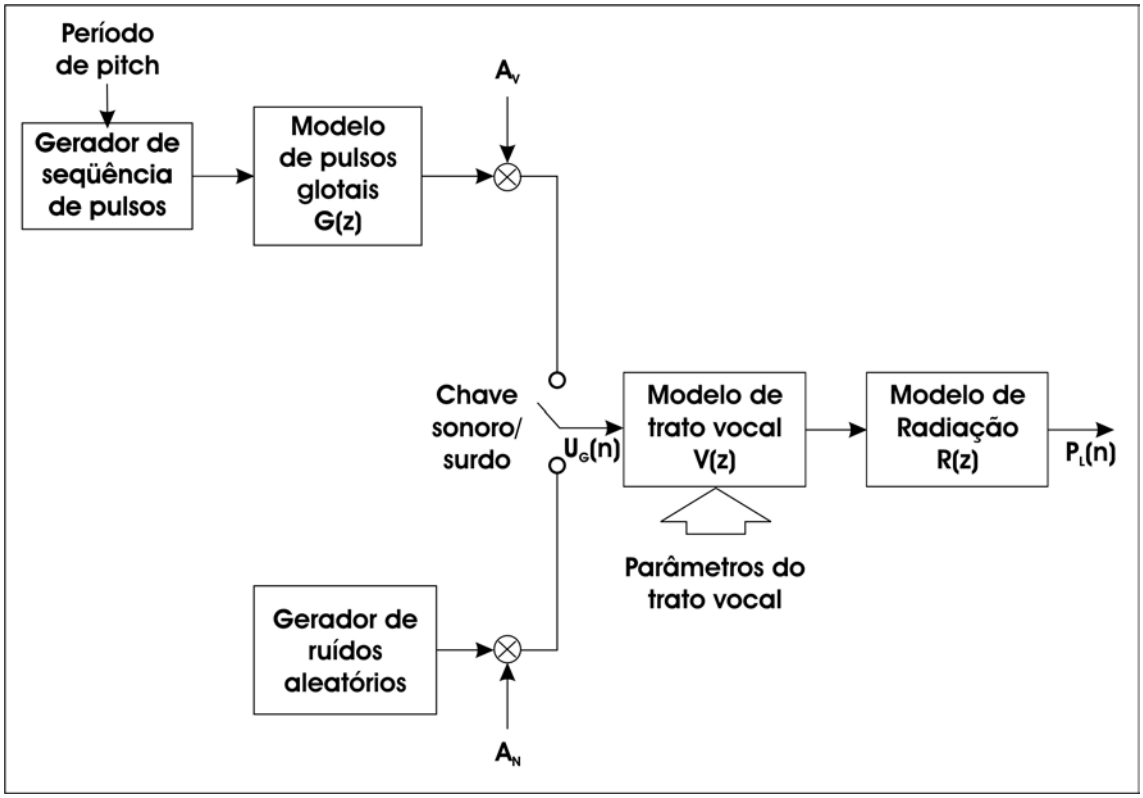

Figura 2.11 – Modelo geral para a geração da fala discreta no tempo [16].

Com o chaveamento entre o gerador de sons sonoros e surdos escolhe-se o modo de excitação. O trato vocal pode ser modelado de várias maneiras diferentes. Em alguns casos é conveniente agrupar os modelos do pulso glotal e a radiação em um único sistema. É conveniente combinar os componentes pulso glotal, radiação e trato vocal, todos juntos e representá-los como uma função de transferência do tipo "all pole",

$$H(z) = G(z).V(z).R(z) \tag{2.4}$$

Neste ponto, as limitações deste modelo é uma questão natural. As deficiências deste modelo não limitam severamente sua aplicabilidade. Primeiro, existe a questão da variação dos parâmetros com o tempo. Em sons contínuos, tais como as vogais, os parâmetros mudam muito lentamente e o modelo trabalha muito bem. Com os sons transientes tais como as consoantes oclusivas o modelo não é muito bom, mas ainda é adequado. É necessário enfatizar que o uso da função de transferência e as funções de resposta em freqüência supõem que é possível representar o sinal da fala em uma base de tempo curta. Os parâmetros do modelo são assumidos constantes sobre os intervalos de tempo tipicamente de 10 a 20 milissegundos. Então a função de transferência V(z) serve para definir estruturas de um modelo cujos parâmetros variam lentamente com o tempo. A segunda limitação é a falta de disposição dos zeros como é requerido teoricamente para os sons nasais e fricativos. Essa é uma limitação para os sons nasais, mas não é muito severa para as fricativas. Zeros podem ser

incluídos no modelo desejado. Terceira, a simples dicotomia da excitação sonora-surda não é adequada para os sons fricativos. Finalmente, o modelo requer que os pulsos glotais sejam espaçados por um inteiro múltiplo do período de amostragem T. Winham e Steiglitz [16] têm considerado as maneiras de eliminar esta limitação nas situações onde é necessário um controle de *pitch* preciso.

2.5 Técnicas de Processamento Digital de Sinais

Nesta seção são apresentados os métodos e as técnicas de processamento de sinais utilizados neste trabalho. Na análise e síntese do sinal da fala foi utilizada a técnica de Codificação da Fala por Predição Linear "*Linear Predictive Speech Coding*" (LPC). Sendo a análise LP - *"Linear Prediction Analysis" (LP)* utilizada na estimação dos coeficientes de predição linear (coeficientes LPC ou vetores LPC) e no cálculo do sinal residual, e a síntese LP - *"Linear Prediction Synthesis" (LP)* utilizada na reconstrução do sinal da fala. Também são apresentados os processos que auxiliam a aplicação da técnica LPC, como: a conversão dos coeficientes LPC para coeficientes LSF "*Line Spectral Frequency*" *(LSF)*; a conversão dos coeficientes LSF para coeficientes LPC; a expansão da largura de faixa; e a pré-ênfase e de-ênfase.

2.5.1 Codificação da fala por predição linear

A codificação da fala pela predição linear "*Linear Predictive Speech Coding*" (LPC) é uma das técnicas de codificação mais populares para os sinais da fala. Ela é uma técnica utilizada em uma variedade de tipos diferentes de codificadores da fala [18]. O método de análise por predição linear é muito utilizado para estimar os parâmetros básicos da fala, como a freqüência fundamental (*pitch*), os formantes, o espectro, a função área do trato vocal e para representar a fala em uma baixa taxa de bits na transmissão ou no armazenamento. A importância deste método está na habilidade do cálculo dos parâmetros estimados da fala com extrema precisão e na relativa velocidade de cálculo.

Aqui é descrita, basicamente, a análise no domínio do tempo [16, 19].

2.5.1.1 Princípios básicos da análise de predição linear

Neste capítulo, seção 2.4, foi descrito o modelo discreto no tempo para a produção da fala. O modelo será repetido aqui com algumas simplificações. A forma particular deste modelo,

apropriada para análise da predição linear, é representada na Figura 2.12. Neste caso, a composição dos efeitos no espectro da radiação, trato vocal e na excitação glotal são representados por um filtro digital variante no tempo onde a função do sistema no estado estacionário é da forma,

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k \, z^{-k}} \tag{2.5}$$

chamado de modelo "all-pole".

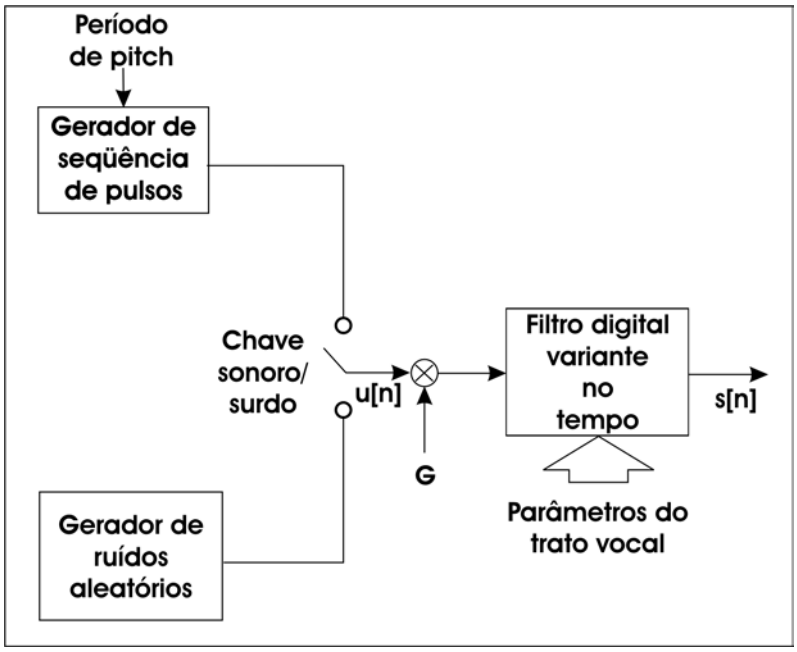

Figura 2.12 – Diagrama de blocos simplificado para o modelo de produção da voz [16].

O sistema da Figura 2.12 é excitado por um trem de impulsos para os sons sonoros e por uma seqüência de ruído aleatório para os sons surdos. Assim os parâmetros deste modelo são: classificação sonoro/surdo, período de *pitch* para sons sonoros, parâmetro do ganho G e os coeficientes $\{a_k\}$ do filtro digital. Estes parâmetros variam lentamente no tempo e são utilizados para identificar um som particular. Em adição a identificação do som que pode ser obtida do espectro, tais informações derivadas podem ser usadas no reconhecimento automático da fala ou nos sistemas de compressão da fala.

A idéia básica em torno da análise preditiva linear é que uma amostra da fala pode ser aproximada como uma combinação linear das amostras anteriores. Pela minimização da soma das diferenças, sobre um intervalo finito, entre a amostra atual de voz e a mesma amostra predita linearmente, pode ser determinado um único conjunto de coeficientes preditores. Os coeficientes preditores são os coeficientes usados na combinação linear.

Como foi visto na seção 2.4 deste capítulo, o modelo simplificado "*all-pole*" é uma representação natural dos sons não nasais. Para os sons nasais e fricativos, de acordo com a teoria acústica, deve-se incluir pólos e zeros na função de transferência. Mas se a ordem p é suficientemente alta, o modelo "*all-pole*" produz uma boa representação para quase todos os sons do aparelho vocal. A maior vantagem deste modelo é que o parâmetro G e os coeficientes $\{a_k\}$ do filtro podem ser estimados em uma maneira direta e de computação eficiente, pelo método de predição linear.

No sistema da Figura 2.12, o modelo "*all-pole*", o sinal de voz s(n) é uma combinação linear dos valores atrasados e alguma entrada u(n) que são relacionados pela equação diferença:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) . \tag{2.6}$$

Um preditor linear com coeficientes de predição α_k é definido como um sistema cuja saída é

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{2.7}$$

onde $\tilde{s}(n)$ é visto como o valor aproximado de $s(n)$ ou o valor predito de $s(n)$.

A função do sistema de p-ésima ordem do preditor linear é o polinômio

$$P(z) = \sum \alpha_k z^{-k} . \tag{2.8}$$

O erro de predição, $e(n)$, é definido como a diferença entre o valor atual $s(n)$ e o valor predito $\tilde{s}(n)$

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) . \tag{2.9}$$

Da Equação (2.9) pode ser visto que a seqüência do erro de predição é a saída de um sistema cuja função de transferência é:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} . \tag{2.10}$$

Isto pode ser visto por comparação das Equações (2.6) e (2.9) que se o sinal da fala obedece o modelo da Equação (2.6) exatamente, e se $\alpha_k = a_k$, então $e(n) = Gu(n)$. Assim, o filtro do erro de predição, $A(z)$, será um *filtro inverso* para o sistema, $H(z)$, da Equação (2.5), isto é,

$$H(z) = \frac{G}{A(z)}. \tag{2.11}$$

O problema básico da análise da predição linear é determinar o conjunto de coeficientes preditores $\{\alpha_k\}$ diretamente do sinal de voz, de maneira a obter uma boa estimação das propriedades espectrais do sinal de voz utilizando-se a Equação (2.11). Os coeficientes do preditor devem ser calculados em curtos segmentos do sinal de voz, ou curtos intervalos de tempo, devido a natureza variante do sinal de voz no tempo. A idéia básica é encontrar o conjunto de coeficientes do preditor que minimiza o erro quadrático médio sobre um curto segmento do sinal de voz. Os parâmetros resultantes são então assumidos como os parâmetros da função do sistema $H(z)$, no modelo de produção de voz.

No método dos mínimos quadrados os parâmetros α_k são obtidos como resultado da minimização do erro quadrático médio ou quadrático total relativo a cada um dos parâmetros.

O erro quadrático total é definido por *E*, onde:

$$E = \sum_n e^2(n) = \sum_n \left[s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right]^2 . \tag{2.12}$$

Minimizando E sem especificar a faixa da somatória, e fazendo:

$$\frac{\partial E}{\partial \alpha_i} = 0, \qquad 1 \le i \le p \tag{2.13}$$

derivando o erro em relação aos coeficientes α_k, obtém-se das Equações (2.12) e (2.13)

$$\frac{\partial E}{\partial \alpha_i} = 2\sum_n \left\{ \left[s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right] \left[\frac{\partial}{\partial \alpha_i} \left(\sum_{k=1}^{p} [-\alpha_i s(n-k)] \right) \right] \right\} = 0$$

$$2\sum_n \left\{ \left[s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right] [-s(n-i)] \right\} = 0$$

$$\sum_n \left\{ -s(n)s(n-i) + \sum_{k=1}^{p} \alpha_k s(n-k)s(n-i) \right\} = 0$$

$$\sum_{k=1}^{p} \alpha_k \sum_n s(n-k)s(n-i) = \sum_n s(n)s(n-i), \qquad 1 \le i \le p\,. \tag{2.14}$$

Para um sinal $s(n)$ definido as Equações (2.14) formam um conjunto de p equações e p incógnitas que podem ser resolvidas para os coeficientes do preditor $\{\alpha_k, 1 \le k \le p\}$ que minimizam E na Equação (2.12).

O erro total mínimo, denotado por E_p é obtido expandindo a Equação (2.12) e substituindo as Equações (2.14). O resultado pode ser mostrado como:

$$E_p = \sum_n s^2(n) - \sum_{k=1}^{p} \alpha_k \sum_n s(n)s(n-k)\,. \tag{2.15}$$

Especificando os limites da somatória sobre n nas Equações (2.12), (2.14) e (2.15) procura-se neste caso determinar um método para calcular os parâmetros, o método da autocorrelação.

2.5.1.1.1 Método da autocorrelação

Assumindo que o erro na Equação (2.12) é minimizado sobre um intervalo de duração infinita, $-\infty < n < \infty$ as Equações (2.14) e (2.15) são reduzidas para:

$$\sum_{k=1}^{p} \alpha_k R(i-k) = R(i), \quad 1 \le i \le p \tag{2.16}$$

$$E_p = R(0) - \sum_{k=1}^{p} \alpha_k R(k) \tag{2.17}$$

onde

$$R(i) = \sum_{n=-\infty}^{\infty} s(n)s(n-i) \tag{2.18}$$

é a função autocorrelação do sinal $s(n)$. Nota-se que R(i) é uma função par de i, isto é:

$$R(-i) = R(i). \tag{2.19}$$

Desde que os coeficientes $R(i-k)$ formam o que é conhecido freqüentemente como a matriz autocorrelação, este método é chamado de método da autocorrelação.

Na Equação (2.17) nota-se que o erro mínimo total consiste de um componente fixo e outro que depende dos coeficientes do preditor.

Na prática o sinal $s(n)$ é conhecido somente em um intervalo finito. Um método comum é multiplicar o sinal $s(n)$ por uma função janela $w(n)$ para obter outro sinal $s'(n)$ que é nulo fora do intervalo $0 \le n \le N-1$:

$$s'(n) = \begin{cases} s(n)w(n), & 0 \le n \le N-1 \\ 0, & outros. \end{cases} \tag{2.20}$$

Neste trabalho utilizou-se a janela de Hamming[(2.1)] [16] que é obtida por:

$$w(n) = \begin{cases} 0{,}54 - 0{,}46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0 & outros \end{cases} \tag{2.21}$$

A autocorrelação é obtida então por:

$$R(i) = \sum_{n=0}^{N-1-i} s'(n)s'(n+i), \qquad i \ge 0. \tag{2.22}$$

2.5.1.1.2 Cálculo do ganho

Fazendo suposições apropriadas sobre o sinal de excitação para o sistema LPC pode-se determinar o ganho G igualando a energia do sinal com a energia do sinal das amostras linearmente preditas.

A Equação (2.9) pode ser reescrita como:

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + e(n). \tag{2.23}$$

(2.1) As janelas mais utilizadas no processamento de sinais da fala são as janelas de Hamming e de Hanning. As propriedades espectrais dessas janelas mostram que elas têm lóbulos principais praticamente idênticos. O que define a escolha entre uma e outra são suas características devido aos lóbulos laterais. Apesar da janela de Hanning apresentar no geral lóbulos laterais com uma maior atenuação que a janela de Hamming, o primeiro lóbulo lateral da janela de Hamming tem uma melhor atenuação que o lóbulo correspondente na janela de Hanning. Devido a este fato a janela de Hamming produz uma melhor resolução ou discriminação em freqüência que a janela de Hanning, sendo então a janela de Hamming a mais utilizada nos processamentos da fala [17, 69].

Comparando as Equações (2.6) e (2.23) observa-se que, para o caso onde $a_k = \alpha_k$, ou seja, os coeficientes atuais do preditor e aqueles do modelo são idênticos, o único sinal de entrada $u(n)$ que resulta no sinal $s(n)$ como saída é $Gu(n) = e(n)$. Assim, o sinal de entrada é proporcional ao erro. Para qualquer outra entrada $u(n)$, a saída do filtro $H(z)$ será diferente de $s(n)$. Entretanto, insistindo que, para qualquer entrada $u(n)$ a energia do sinal de saída precisa ser igual à do sinal $s(n)$, pode-se pelo menos especificar a energia total no sinal de entrada. Desde que o filtro $H(z)$ é fixo, fica claro do exposto acima que a energia do sinal de entrada $Gu(n)$ precisa ser igual à energia total no sinal erro, o qual é obtido por *Ep* na Equação (2.17).

Neste ponto é necessário fazer algumas suposições sobre $u(n)$ para relacionar G com as quantidades conhecidas, isto é, com os α_k's e os coeficientes da autocorrelação. Existem dois casos de interesse para a excitação: para sons os sonoros e para os sons surdos.

a) Para os sons sonoros

É razoável assumir $u(n) = \delta(n)$, isto é, a excitação é o impulso unitário em n = 0. A saída é então a resposta ao impulso h(n), onde:

$$h(n) = \sum \alpha_k h(n-k) + G\delta(n) . \tag{2.24}$$

A autocorrelação $\hat{R}(i)$ da resposta ao impulso $h(n)$ tem um relacionamento interessante com a autocorrelação $R(i)$ do sinal $s(n)$. Multiplicando a Equação (2.24) por $h(n\text{-}i)$ e fazendo a somatória sobre todo n, resulta:

$$\hat{R}(i) = \sum_{k=1}^{p} \alpha_k \hat{R}(i-k), \quad 1 \le |i| \le \infty \tag{2.25}$$

$$\hat{R}(0) = \sum_{k=1}^{p} \alpha_k \hat{R}(k) + G^2 . \tag{2.26}$$

Dada a condição de que a energia total em $h(n)$ e $s(n)$ devem ser iguais, resulta que

$$\hat{R}(0) = R(0) \tag{2.27}$$

desde que o coeficiente de ordem zero é igual a energia total no sinal. Da Equação (2.27) e da semelhança entre as Equações (2.16) e (2.17), conclui-se que:

$$\hat{R}(i) = R(i), \qquad 0 \le i \le p. \tag{2.28}$$

Esta equação diz que os primeiros p+1 coeficientes da autocorrelação da resposta ao impulso de $H(z)$ são idênticos aos correspondentes coeficientes da autocorrelação do sinal.

Das Equações (2.26), (2.28) e (2.17), o ganho é igual a:

$$G^2 = E_p = R(0) - \sum_{k=1}^{p} \alpha_k R(k) \tag{2.29}$$

onde G^2 é a energia total da entrada $G\delta(n)$.

b) Para os sons surdos

Neste caso é mais razoável supor que a entrada $u(n)$ é um ruído branco com média zero e variância unitária, isto é:

$$E[u(n)] = 0 \text{ para todo n e } E[u(n)u(n-i)] = \delta(i). \tag{2.30}$$

Se o sistema for excitado com uma entrada randômica $Gu(n)$ e a saída for chamada de $g(n)$, então:

$$g(n) = \sum_{k=1}^{p} \alpha_k g(n-k) + Gu(n). \tag{2.31}$$

Chamando de $\tilde{R}(i)$ a autocorrelação da função $g(n)$ então:

$$\tilde{R}(i) = E[g(n)g(n-i)] = \sum_{k=1}^{p} \alpha_k E[g(n-k)g(n-i)] + E[Gu(n)g(n-i)]$$

$$= \sum_{k=1}^{p} \alpha_k \tilde{R}(i-k), i \neq 0 \tag{2.32}$$

desde que $E[u(n)g(n-i)] = 0$ para i > 0, porque $u(n)$ não é correlacionado com outro sinal anterior a $u(n)$. Para i = 0, tem-se:

$$\tilde{R}(0) = \sum_{k=1}^{p} \alpha_k \tilde{R}(k) + GE[u(n)g(n)]$$

$$= \sum_{k=1}^{p} \alpha_k \tilde{R}(k) + G^2 \tag{2.33}$$

desde que $E[u(n)g(n)] = E[u(n)(G(n) + termos \quad anteriores \quad a \quad n)] = G$. Desde que a energia na resposta para $Gu(n)$ deve ser igual a energia do sinal, então

$$\tilde{R}(i) = R(i), \qquad 0 \le i \le p \tag{2.34}$$

ou

$$G^2 = R(0) - \sum_{k=1}^{p} \alpha_k R(k) \tag{2.35}$$

como no caso para sons sonoros.

2.5.1.1.3 Cálculo dos parâmetros do preditor

Os coeficientes do preditor α_k, $1 \le k \le p$ são calculados a partir das Equações (2.16) que compõem um conjunto de p equações com p incógnitas. Existem vários métodos padrões para calcular estas equações, como redução de Gauss ou método de eliminação e o método da redução de Crout. Esses métodos gerais requerem $p^3/3 + O(p^2)$ operações[(2.2)] (multiplicações e divisões) e p^2 locais de memória.

Devido a forma especial das Equações (2.16) é possível reduzir o tempo de computação e o armazenamento em sua resolução. Estas equações podem ser representadas na forma de matriz como:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \cdots \\ \cdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \cdots \\ \cdots \\ R(p) \end{bmatrix}. \tag{2.36}$$

Nota-se que a matriz p x p de valores de autocorrelação é uma matriz Toeplitz. Isto é, simétrica e os elementos ao longo de qualquer diagonal são iguais. Levinson derivou um procedimento recursivo elegante para resolver este tipo de equação. O procedimento foi posteriormente reformulado. O método de Levinson assume que o vetor coluna do lado direito da Equação (2.36) é um vetor coluna geral. Usando o fato de que esse vetor coluna contém os

[(2.2)] A notação "O(.)" denota "na ordem de" e indica a aproximação.

mesmos elementos encontrados na matriz de autocorrelação, outro método atribuído a Durbin torna o método duas vezes mais rápido do que o algoritmo de Levinson. O método requer somente 2p localizações de memória $p^2 + O(p)$ operações[(2.1)].

O ***método recursivo de Durbin*** pode ser especificado como segue:

$$E_0 = R(0) \tag{2.37a}$$

$$k_i = \left[R(i) + \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j) \right] / E_{i-1} \tag{2.37b}$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, \quad 1 \le j \le i-1 \tag{2.37c}$$

$$E_i = (1 - k_i^2) E_{i-1}. \tag{2.37d}$$

As Equações (2.37b) a (2.37d) são resolvidas recursivamente para i = 1,2,..., p. A solução final é obtida por:

$$\alpha_j = \alpha_j^{(p)}, \qquad 1 \le j \le p. \tag{2.37e}$$

Nota-se que na obtenção da solução para um preditor de ordem p, são realmente computadas as soluções para todos os preditores de ordem inferiores a p.

Para a maioria das aplicações a solução da Equação (2.16) não constitui o maior esforço computacional. O cálculo dos coeficientes da autocorrelação requer pN operações que, como é freqüentemente o caso, pode ser dominante no tempo de computação se N >> p.

A solução da Equação (2.36) não é afetada se todos os coeficientes da autocorrelação são divididos por uma constante. Em particular, se todos $R(i)$ são normalizados pela divisão por $R(0)$ encontra-se os coeficientes normalizados da autocorrelação $r(i)$:

$$r(i) = \frac{R(i)}{R(0)} \tag{2.38}$$

os quais têm a propriedade de que $|r(i)| \le 1, \quad 0 \le i \le p.$

[(2.1)] A notação "O(.)" denota "na ordem de" e indica a aproximação.

Se os coeficientes da autocorrelação são normalizados como na Equação (2.38), então o erro mínimo E_i é também dividido por $R(0)$. Esta quantidade resultante é chamada de erro normalizado V_i:

$$V_i = \frac{E_i}{R(0)} = 1 - \sum_{k=1}^{i} \alpha_k r(k) \tag{2.39}$$

com $0 \le V_i \le 1, \qquad i \ge 0.$

Das Equações (2.37d) e (2.39), o erro normalizado para i = p é obtido por:

$$V_p = \prod_{i=1}^{p} \left(1 - k_i^2\right) \tag{2.40}$$

onde as quantidades k_i estão na faixa $-1 \le k_i \le 1$. Esta condição dos parâmetros k_i é importante desde que é uma condição necessária e suficiente para que todas as raízes do polinômio $A(z)$ estejam dentro do círculo unitário, garantindo assim a estabilidade do sistema $H(z)$.

É preciso ser notado que esta garantia teórica de estabilidade para o método da autocorrelação pode não ser válida na prática se a função autocorrelação é calculada sem precisão suficiente. Markel e Gray mostraram que esses efeitos indesejáveis podem ser minimizados pela pré-enfatização do sinal de voz para tornar o espectro tão plano quanto possível [16].

O algoritmo de Durbin permite um teste conveniente para a estabilidade desde que é necessário e suficiente que os parâmetros k_i precisam satisfazer a condição:

$$-1 \le k_i \le 1. \tag{2.41}$$

Desta forma, se no processo de determinação dos coeficientes do preditor $\{\alpha_i\}$ alguma das quantidades k_i violar a Equação (2.41) sabe-se que existem raízes de $A(z)$ fora do círculo unitário.

2.5.1.2 Resumo dos procedimentos para aplicação da técnica LP – Análise e síntese LP

Para tornar operacional a aplicação da técnica de predição linear (técnica LP) na técnica de codificação da fala por interpolação de ondas (técnica WI) é realizado um resumo do procedimento em três itens:

- Procedimentos para o *Cálculo dos Coeficientes LPC* – Durante a Análise LP;
- Procedimentos para o *Cálculo do Sinal Residual* - Análise LP (Filtro de Análise LP);
- Procedimentos para a *Síntese do Sinal por Predição Linear* – Síntese LP (Filtro de Análise LP).

Procedimentos para o Cálculo dos Coeficientes LPC – Durante a Análise LP

Seja s[n] o sinal da fala que é dividido em quadros consecutivos $\ldots, (s[n]_1), (s[n]_2)\ldots, (s[n]_{l-1}), (s[n]_l), (s[n]_{l+1})\ldots$ cada um com comprimento de L_q amostras, onde $(s[n]_l)$ representa as amostras do quadro atual, $(s[n]_{l-1})$ as amostras do quadro anterior (ou passado) e $(s[n]_{l+1})$ as amostras do quadro posterior ou em avanço. Os coeficientes LPC, $\vec{\alpha} = \{\alpha(1),\alpha(2),\cdots,\alpha(p)\}$, correspondentes aos quadros do sinal de voz, $\ldots, (s[n]_1), (s[n]_2)\ldots, (s[n]_{l-1}), (s[n]_l), (s[n]_{l+1})\ldots$, podem ser encontrados executando os seguintes passos:

- Inicialmente o conjunto de amostras relacionadas com a análise do quadro atual $(s[n]_l)$, é selecionado sobre o sinal da fala s[n] por uma janela de análise, $(s[n])_{Lwl}$, com comprimento L_{wl} amostras. Em geral $(s[n])_{Lwl}$ não é totalmente coincidente com $(s[n]_l)$ podendo ter seu centro coincidente ou não com o centro de $(s[n]_l)$ e englobando ou não as amostras do quadro em atraso $(s[n]_{l-1})$ ou do quadro em avanço $(s[n]_{l+1})$, conforme o objetivos e compromisso da análise. Também as janelas de análise consecutivas têm M amostras em sobreposição o que garante uma suavização entre os parâmetros estimados.

- Aplica-se as Equações (2.20) e (2.21) passando a seqüência das amostras selecionadas da janela de análise $(s[n])_{Lwl}$ pela janela de Hamming obtendo-se a correspondente seqüência $s'[n]$. Aqui essas equações são reescritas como:

$$s'[n] = \begin{cases} (s[n])_{Lwl} \cdot w[n], & 0 \le n \le N-1 \\ 0, & outros. \end{cases} \tag{2.42}$$

e a janela de Hamming obtida por:

$$w[n] = \begin{cases} 0{,}54 - 0{,}46\cos\left(\dfrac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & outros. \end{cases} \tag{2.43}$$

- Aplica-se as Equações (2.22) para calcular os coeficientes da autocorrelação $R(i)$ que compõem o vetor de coeficientes da autocorrelação $\vec{R} = \{R(0),R(1),R(2),\cdots,R(p)\}$. Aqui a Equação (2.22) é reescrita como:

$$R(i) = \sum_{n=0}^{N-1-i} s'(n)s'(n+i), i \geq 0. \tag{2.44}$$

- Aplica-se a Equação (2.38) para normalizar os coeficientes da autocorrelação $R(i)$ obtendo-se os coeficientes correspondentes normalizados $r(i)$ que compõem o vetor de coeficientes da autocorrelação normalizados $\vec{r} = \{r(0),r(1),r(2),\cdots,r(p)\}$. Aqui a Equação (2.38) é reescrita como:

$$r(i) = \frac{R(i)}{R(0)} \tag{2.45}$$

- Resolve-se o sistema de Equações (2.16) usando o algoritmo de Durbin utilizando-se as Equações (2.37), obtendo finalmente o vetor de coeficientes LPC $\vec{\alpha} = \{\alpha(1),\alpha(2),\cdots,\alpha(p)\}$ correspondente a seqüência de amostras na janela de análise $(s[n])_{Lwl}$ em correspondência a $(s[n]_l)$, quadro atual do sinal.

- Estes procedimento são executados para cada janela de análise em correspondência a cada um dos quadros …, $(s[n]_1)$, $(s[n]_2)$…, $(s[n]_{l-1})$, $(s[n]_l)$, $(s[n]_{l+1})$… .

Procedimentos para Análise do Sinal por Predição Linear – Análise LP (O Filtro de Análise LP)

Reescrevendo a Equação (2.9), $e(n) = s(n) - \tilde{s}(n)$ torna-se:

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{2.46}$$

que é utilizada na análise do sinal.

Aplicando-se a transformada-z chega-se a $E(z) = S(z) - \sum_{k=1}^{p} \alpha_k S(z) z^{-k}$ que pode ser escrita como:

$$E(z) = S(z)\left[1 - \sum_{k=1}^{p} \alpha_k z^{-k}\right] \tag{2.47}$$

onde $A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}$ foi definido como *filtro do erro de predição* pela Equação (2.10). Esse filtro também é chamado de *filtro de análise LP.* A Equação (2.47) então torna-se:

$$E(z) = S(z)A(z) \tag{2.48}$$

e pode ser interpretada como um *sistema para análise do sinal por predição linear*, ou seja, a seqüência $e(n)$ do erro de predição (ou sinal residual) é a saída de um sistema cuja função de transferência é obtida por $A(z)$ sendo a seqüência $s(n)$ do sinal da fala a entrada do sistema. O *sistema para análise por predição linear* é mostrado na Figura 2.13.

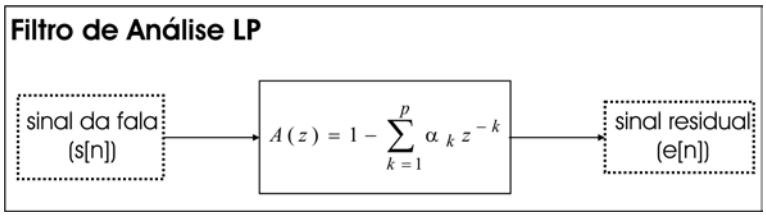

Figura 2.13 – Diagrama esquemático do Filtro de Análise LP.

Assim, considerando que o vetor de coeficientes LPC $\vec{\alpha} = \{\alpha(1), \alpha(2), \cdots, \alpha(p)\}$ correspondente ao quadro atual $(s[n]_l)$ já foi calculado, e que as últimas p amostras do quadro anterior $(s[n]_{l-1})$ mais as amostras do quadro atual $(s[n]_l)$ estão disponíveis, os procedimentos para a análise LP do sinal do quadro atual são:

- aplica-se a Equação (2.46) que torna-se $(e[n]_l) = (s[n]_l) - \sum_{k=1}^{p} \alpha_k (s[n-k]_l)$ e calcula-se a seqüência de amostras de $(e[n]_l)$.

- O procedimento anterior é executado para cada quadro dado por $\ldots, (s[n]_1), (s[n]_2) \ldots, (s[n]_{l-1}), (s[n]_l), (s[n]_{l+1}) \ldots$ obtendo-se em correspondência a cada um deles o sinal residual dado por $\ldots, (e[n]_1), (e[n]_2) \ldots, (e[n]_{l-1}), (e[n]_l), (e[n]_{l+1}) \ldots$.

Procedimentos para Síntese do Sinal por Predição Linear – Síntese LP (O Filtro de Síntese LP)

A Equação (2.9) também pode ser escrita como:

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + e(n) \tag{2.49}$$

que é utilizada na síntese do sinal da fala.

A Equação (2.48) $E(z) = S(z)A(z)$ pode ser escrita como:

$$S(z) = E(z) \cdot \left[\frac{1}{A(z)} \right] \tag{2.50}$$

e pode ser interpretada como um sistema para síntese do sinal por predição linear. Assim, a seqüência $s(n)$ do sinal da fala sintetizado é a saída de um sistema cuja função de transferência é obtida por $1/A(z)$, sendo a seqüência $e(n)$ do erro de predição (ou sinal residual) a entrada do sistema. Aqui a função de transferência $1/A(z)$ é definida como filtro de síntese *LP*. O sistema para síntese por predição linear é mostrado na Figura 2.14.

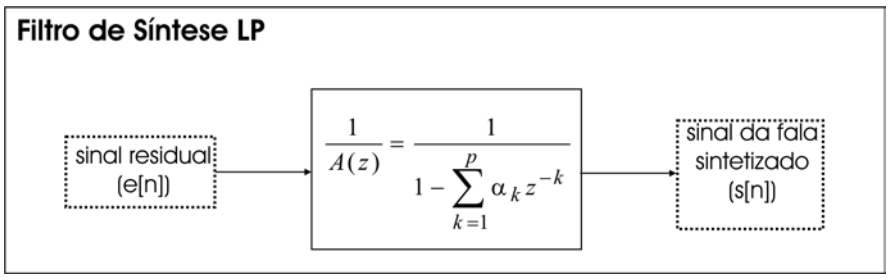

Figura 2.14 – Diagrama esquemático do Filtro de síntese LP.

Assim, considerando que o vetor de coeficientes LPC $\vec{\alpha} = \{\alpha(1), \alpha(2), \cdots, \alpha(p)\}$ correspondente ao quadro atual $(s[n]_l)$ está disponível, e que as últimas p amostras do quadro anterior $(s[n]_{l-1})$ do sinal sintetizado, mais as amostras do sinal residual $(e[n]_l)$, correspondentes ao quadro atual $(s[n]_l)$ estão disponíveis, os procedimentos para a síntese LP do sinal do quadro atual são:

- aplica-se a Equação (2.49) que torna-se $(s[n]_l) = \sum_{k=1}^{p} \alpha_k (s[n-k]_l) + (e[n]_l)$ e calcula-se a seqüência de amostras de $(s[n]_l)$.

- O procedimento anterior é executado para cada quadro do sinal residual obtido por $\ldots, (e[n]_1), (e[n]_2) \ldots, (e[n]_{l-1}), (e[n]_l), (e[n]_{l+1}) \ldots$ obtendo-se em correspondência a cada um deles o sinal da fala sintetizado obtido por $\ldots, (s[n]_1), (s[n]_2) \ldots, (s[n]_{l-1}), (s[n]_l), (s[n]_{l+1}) \ldots$.

2.5.1.3 Representação dos coeficientes de predição linear

Para uma transmissão eficiente das informações que os coeficientes de predição linear, coeficientes LPC, carregam, em geral, eles precisam ser quantizados e interpolados. A interpolação possibilita uma transmissão das informações sobre os coeficientes LPC com uma freqüência menor (isto é, a uma taxa baixa, por exemplo, de uma vez por quadro), desta forma reduzindo a taxa de bits [2].

Entretanto, a quantização e a interpolação direta dos coeficientes LPC é problemática porque pequenas mudanças nos coeficientes LPC podem resultar em grandes mudanças no espectro de potência, e possivelmente, em filtros de síntese LP instáveis. Assim a interpolação e a quantização são geralmente realizadas em versões transformadas dos coeficientes LPC. Um número de transformações biunívocas para os coeficientes LPC tem sido desenvolvido para minimizar esses problemas. Entre elas os coeficientes PARCOR, os coeficientes Logaritmo da Relação de Áreas, (LAR), e os coeficientes "*Line Spectral Frequencies*" (LSF) [10]**.**

A representação mais utilizada para os coeficientes LPC é através dos coeficientes LSF's também chamados de "*Line Spectral Pairs*" (LSP). Os LSF's provêem não somente a estabilidade dos coeficientes interpolados, mas também facilitam a manipulação espectral e possuem propriedades adequadas à quantização [14].

Durante a análise LP, em geral, os coeficientes LPC são estimados uma vez por quadro e transformados para LSF's. No domínio LSF são realizadas as quantizações e as interpolações necessárias. Para evitar mudanças bruscas que ocorrem entre os coeficientes de quadros sucessivos, eles são interpolados por sub-quadros, no domínio LSF, de forma que eles evoluem de forma suave ao longo do quadro. Os LSF's são então transformados de volta para os coeficientes LPC para configurar o filtro $A(z)$, o filtro de análise LP para o cálculo do sinal residual, ou o filtro $1/A(z)$, filtro de síntese LP para a síntese do sinal da fala.

2.5.1.3.1 Cálculo dos coeficientes LSF's (Conversão dos LPC's para LSF's)

A transformação dos coeficientes de predição linear em coeficientes LSF foi introduzida por Itakura em 1975 [10], tendo suas propriedades sido estudadas mais tarde por Soong e Juan em 1984 [20]. Por definição, os coeficientes LSF são as freqüências correspondentes às raízes de dois polinômios de ordem $p+1$, $P(z)$ e $Q(z)$, derivados do filtro de análise LP $A(z)$, de ordem p. $P(z)$ corresponde ao trato vocal com a fonte glotal completamente fechada (coeficiente de reflexão $k_{p+1}=1$) e Q(z) representa o trato vocal com a fonte glotal completamente aberta (coeficiente de reflexão $k_{p+1}=-1$). Através da recursão no cálculo dos coeficientes de predição linear para uma ordem superior,

$$A_{p+1}(z) = A_p(z) + k_{p+1} z^{-(p+1)} A_p(z^{-1})$$

os polinômios $P(z)$ (simétrico) e $Q(z)$ (anti-simétrico) são obtidos por [21]:

$$P(z) = A_p(z) + z^{-(p+1)} A(z^{-1}) \tag{2.51}$$

e $$Q(z) = A_p(z) - z^{-(p+1)} A(z^{-1}), \tag{2.52}$$

onde: $$A(z) = \frac{P(z) + Q(z)}{2} \quad . \tag{2.53}$$

Existem três propriedades importantes de $P(z)$ e $Q(z)$ [20]:

- Todas as raízes dos polinômios $P(z)$ e $Q(z)$ estão sobre o circulo unitário;
- As raízes de $P(z)$ e $Q(z)$ são entrelaçadas;
- A propriedade de fase mínima de $A(z)$ pode ser preservada, se as duas primeiras propriedades são obedecidas após a quantização ou a interpolação.

A partir da primeira propriedade, observa-se que as raízes de $P(z)$ e $Q(z)$ podem ser escritas em termos de ω_i (como $e^{j\omega_i}$). Estas freqüências ω_i (ou posição angular) são chamadas de LSF's.
Os polinômios $P(z)$ e $Q(z)$ sendo simétrico e anti-simétrico respectivamente, têm raízes em $z = 1$ e $z = -1$ que podem ser removidas por divisão polinomial. Assim os polinômios auxiliares $N_1(z)$ e $N_2(z)$ tornam-se

$$N_1(z) = \frac{P(z)}{1 + z^{-1}} \text{ e } N_2(z) = \frac{Q(z)}{1 - z^{-1}} \text{ para p par} \tag{2.54}$$

$$N_1(z) = P(z) \quad \text{e} \quad N_2(z) = \frac{Q(z)}{1 - z^{-2}} \text{ para p ímpar} \tag{2.55}$$

A partir das Equações (2.54) e (2.55) nota-se que $N_1(z)$ e $N_2(z)$ têm ordem par e são simétricos. As raízes ocorrem em pares de complexo conjugados e, portanto somente as raízes na parte superior do círculo precisam ser calculadas. Fazendo as ordens de $N_1(z)$ e $N_2(z)$ serem de $2m$ e $2n$, respectivamente

$$m = \frac{p}{2} \text{ e } n = \frac{p}{2} \text{ para p par,} \tag{2.56}$$

$$m = \frac{p+1}{2} \text{ e } n = \frac{p-1}{2} \text{ para p ímpar.} \tag{2.57}$$

Então mostrando explicitamente a simetria dos coeficientes do polinômios

$$N_1(z) = 1 + n_1(1)z^{-1} + n_1(2)z^{-2} + \cdots + n_1(m)z^{-m} + \cdots + n_1(1)z^{-(2m-1)} + z^{-2m} \tag{2.58}$$

$$N_2(z) = 1 + n_2(1)z^{-1} + n_2(1)z^{-2} + \cdots + n_2(n)z^{-n} + \cdots + n_2(1)z^{-(2n-1)} + z^{-2n}. \tag{2.59}$$

$N_1(z)$ contribui com m pares de raízes conjugadas e $N_2(z)$ contribui com n pares de raízes conjugadas ($m+n=p$). No círculo unitário, o termo de fase linear pode ser removido para obter duas séries de fase zero expandidas em cossenos:

$$N_1(e^{j\omega}) = e^{-j\omega m} N_1'(\omega) \tag{2.60}$$

$$N_2(e^{j\omega}) = e^{-j\omega n} N_2'(\omega) \tag{2.61}$$

onde:

$$N_1'(\omega) = 2\cos m\omega + 2n_1(1)\cos(m-1)\omega + \cdots + 2n_1(m-1)\cos\omega + n_1(m) \tag{2.62}$$

$$N_2'(\omega) = 2\cos n\omega + 2n_2(1)\cos(n-1)\omega + \cdots + 2n_2(n-1)\cos\omega + n_2(n). \tag{2.63}$$

Soong e Juang [20, 22] propuseram um método numérico com cálculo direto utilizando-se a transformada discreta do cosseno para determinar as raízes de $N_1'(\omega)$ e $N_2'(\omega)$ que são os LSF's. Sobre a introdução da transformação cossenoidal da freqüência:

$$x = \cos\omega, \tag{2.64}$$

Kabal e Ramachandran [21] notaram que as Equações (2.62) e (2.63) podiam ser reformuladas em termos dos polinômios de Chebyshev. Assim eles usaram uma expansão dos polinômios de Chebyshev de $m-ésima$ ordem em x:

$$T_m(x) = \cos(m\omega), \tag{2.65}$$

onde $T_m(x) = 2xT_{m-1}(x) + T_{m-2}(x)$. As eqs. (2.62) e (2.63) tornam-se

$$N_1'(x) = 2T_m(x) + 2n_1(1)T_{m-1}(x) + \cdots + 2n_1(m-1)T_1(x) + n_1(m) \tag{2.66}$$

$$N_2'(x) = 2T_n(x) + 2n_2(1)T_{n-1}(x) + \cdots + 2n_2(n-1)T_1(x) + n_2(n)\,. \tag{2.67}$$

As raízes dos polinômios expandidos são determinadas iterativamente pela procura da mudança de sinal na faixa [-1,+1] e então os coeficientes LSF's são calculados pelo uso de:

$$\omega = arc\cos x \tag{2.68}$$

que é a função inversa da Equação (2.64).

2.5.1.3.2 Cálculo dos coeficientes LPC's (Conversão dos LSF's para LPC's)

A conversão dos coeficientes LSF para coeficientes LPC é executada com base na Equação (2.53) reescrita aqui por conveniência:

$$A(z) = \frac{P(z)+Q(z)}{2} = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}\,. \tag{2.69}$$

Soong e Juang [20] mostraram que $P(z)$ e $Q(z)$ podem ser escritos como:

$$P(z) = (1-z^{-1})\prod_{i=1}^{p/2}\left(1 - z^{-1}e^{j\omega_i}\right)\left(1 - z^{-1}e^{-j\omega_i}\right) = (1-z^{-1})\prod_{i=1}^{p/2}\left(1 - 2z^{-1}\cos\omega_i + z^{-2}\right) \tag{2.70}$$

onde ω_i $(1 \le i \le p/2)$ são as freqüências angulares das raízes de $P(z)$ ou coeficientes LSF's do polinômio simétrico e:

$$Q(z) = (1+z^{-1})\prod_{i=1}^{p/2}\left(1 - z^{-1}e^{j\theta_i}\right)\left(1 - z^{-1}e^{-j\theta_i}\right) = (1-z^{-1})\prod_{i=1}^{p/2}\left(1 - 2z^{-1}\cos\theta_i + z^{-2}\right) \tag{2.71}$$

onde θ_i $(1 \le i \le p/2)$ são as freqüências angulares das raízes de $Q(z)$ ou coeficientes LSF's do polinômio anti-simétrico. Como já foi mencionado, estes coeficientes são entrelaçados um como outro no intervalo $(0,\pi)$ da seguinte forma:

$$0 < \omega_1 < \theta_1 < \omega_2 < \theta_2 < \cdots < \omega_{p/2} < \theta_{p/2} < \pi\,. \tag{2.72}$$

Supondo que os coeficientes LSF's, ω_i $(1 \le i \le p/2)$ e θ_i $(1 \le i \le p/2)$ são conhecidos, os coeficientes LPC, α_i, $i = 1, \cdots, p$, podem ser recuperados substituindo-se as Equações (2.70) e (2.71) na Equação (2.69) e separando os termos com as apropriadas potências de z.

2.5.1.3.3 Algoritmos para o Cálculo dos LFS's e Conversão para os LPC's

Os algoritmos utilizados neste trabalho são descritos por Kabal e Ramachandran em [21]:

(i) Algoritmo para calcular os coeficientes LSF a partir dos coeficientes LPC; e

(ii) Algoritmo para converter os coeficientes LSF para coeficientes LPC.

2.5.2 Expansão da Largura de Faixa

A análise LP não estima com precisão o envelope espectral para sons sonoros da fala com alta freqüência do *pitch*. As informações espectrais sobre o sinal periódico estão contidas somente nas harmônicas. Mas para fala com alta freqüência do *pitch* o espaçamento entre as harmônicas é tão grande para prover uma adequada amostragem do envelope espectral que resulta em uma subestimação da largura do formante [23]. Assim o envelope espectral tende a ter alta ressonância (picos afiados) e com estreita largura de faixa [24]. Isto significa que os polos do filtro estão muito próximos do círculo unitário, e então, o filtro fica marginalmente estável. Tal estabilidade marginal no filtro LP pode aumentar as chances de se atingir sobre valores na quantização dos coeficientes LSF que causam chiados na fala quantizada. Uma solução é empregar a expansão da largura de faixa, para expandir a largura de faixa na resposta em freqüência do filtro. Nesta técnica os coeficientes de predição linear α_k são trocados por $\gamma^k \alpha_k \quad (k = 1, \cdots, p)$ onde p é a ordem do preditor linear. Esta multiplicação move todos os pólos do filtro em direção ao centro do círculo afastando-os da circunferência de raio unitário por um fator γ. Isto resulta em picos suavizados e largura de faixa mais larga na resposta em freqüência do filtro tornando-o mais estável. O fator γ denominado de fator de expansão da largura de faixa controla quanto os pólos são movidos para dentro. Se ν_i é o raio do $i-ésimo$ pólo, então a largura de faixa deste pólo é definida por

$$B_i = -\frac{1}{\pi T} \ln(\nu_i) \tag{2.73}$$

onde T é o intervalo de amostragem. A multiplicação do raio por γ expande a largura de faixa para $B_i + \Delta B$, onde:

$$\Delta B = -\frac{1}{\pi T}\ln(\gamma) \tag{2.74}$$

Valores típicos para γ estão entre 0,988 e 0,996 que correspondem a expansão de 10 a 30 Hz na largura de faixa [23].

2.5.3 Pré-ênfase e De-ênfase

O espectro dos sons sonoros da fala normalmente tem uma queda de 6-dB/oitava, que resulta em uma faixa dinâmica espectral alta. De fato, o espectro da fala tem um leve formato de um filtro passa baixa, ou seja, nas freqüências mais altas o sinal fica com uma energia relativa mais baixa. Esta faixa dinâmica espectral alta usualmente resulta em uma aproximação imprecisa dos formantes de maior freqüência e algumas vezes resultam em um mau condicionamento na matriz de autocorrelação, e conseqüentemente afeta a precisão numérica dos coeficientes LP [17]. Para diminuir este efeito o sinal da fala é usualmente pré-enfatizado antes da análise LP. Isto é realizado passando o sinal da fala $s[n]$ por um filtro de pré-ênfase que geralmente é da forma:

$$H_{pre}(z) = 1 - \lambda z^{-1} \tag{2.75}$$

onde $H_{pre}(z)$ é um filtro passa alta "suave" com um único zero em λ. A constante λ, conhecida como *fator de pré-ênfase*, controla o grau de pré-ênfase. Típicos valores de λ são tomados na faixa $0{,}9 \le \lambda \le 1{,}0$ [17]. Passando o sinal $s[n]$ pelo filtro $H_{pre}(z)$ resulta em $s_{pre}[n]$ (no domínio do tempo)

$$s_{pre}[n] = s[n] - \lambda\, s[n-1] \tag{2.76}$$

onde $s_{pre}[n]$ é o sinal da fala pré-enfatizado. Para cancelar os efeitos da pré-ênfase, realiza-se a de-ênfase, ou seja, o sinal enfatizado é passado por um filtro de de-ênfase $H_{deenf}(z)$ definido como o inverso do filtro de pré-ênfase $H_{pre}(z)$. A função de transferência $H_{deenf}(z)$ é:

$$H_{deenf}(z) = \frac{1}{1-\lambda z^{-1}}. \tag{2.77}$$

Passando o sinal $s_{pre}[n]$ pelo filtro $H_{deenf}(z)$ resulta em $s[n]$ (no domínio do tempo)

$$s[n] = s_{pre}[n] + \lambda\, s[n-1]\,. \tag{2.78}$$

2.6 Considerações Finais Deste Capítulo

Neste capítulo foram apresentados uma breve visão do processamento digital de sinais da fala, o processo de produção da fala e um modelo matemático discreto no tempo para a produção dos sinais de voz. Este modelo que pode ser obtido a partir da aplicação da física acústica no processo de produção da fala serve de base para a aplicação das técnicas de processamento digital de sinais. Também foi apresentada a técnica de Codificação da Fala por Predição Linear "*Linear Predictive Speech Coding*" (LPC) que é utilizada na estimação dos coeficientes de predição linear (coeficientes LPC ou vetores LPC) e no cálculo do sinal residual – durante a *análise LP*, e na reconstrução do sinal da fala – durante a *síntese LP*. Também foram apresentados alguns processos auxiliares para a aplicação da técnica LPC.

Estas técnicas são utilizadas como processos componentes e auxiliares para a técnica principal deste trabalho, denominada de codificação da fala por interpolação de ondas – "*Waveform Interpolation Speech Coding*", técnica WI, que é descrita no capítulo 4.

Capítulo 3

CODIFICAÇÃO DA FALA - UMA VISÃO GERAL

3.1 Introdução

A representação digital de sinais da fala de modo eficiente tem se tornado uma área de grande importância. Quando um sinal é transmitido, o número de bits que representa cada segundo da fala, ou seja, a taxa de bits produzida é um parâmetro importante na definição da largura de faixa do canal de transmissão. Da mesma forma, o advento das tecnologias multimídia e a necessidade de armazenamento de grandes quantidades de informação para utilização posterior exigem a necessidade de reduzir a taxa de bits, já que ela determina o espaço requerido na unidade de armazenamento. A redução da taxa de bits na representação dos sinais da fala sem comprometimento da qualidade é o principal objetivo da codificação.

Há poucas décadas, várias técnicas têm sido propostas, analisadas e desenvolvidas. Algumas usam taxas de bits elevadas e obtêm facilmente alta qualidade da fala, enquanto outras usam taxas menores, a custa de uma degradação da qualidade da fala.

Neste capítulo será feita uma breve discussão sobre as técnicas que são utilizadas atualmente, e sobre aquelas que poderão ser utilizadas no futuro. Tradicionalmente, os codificadores da fala são divididos em duas classes – *codificadores de forma de onda* e *codificadores da fonte*, também conhecidos como codificadores paramétricos, os "*voice coders"* (vocoders). Os *codificadores de forma de onda* conseguem produzir uma fala de alta qualidade, mas a custa de taxas de bits elevadas. Já os codificadores da fonte são usados em taxas de bits muito baixas, mas tendem a produzir uma fala de qualidade sintética. Mais recentemente foi introduzido uma nova classe de codificadores, chamados de codificadores híbridos, os quais usam ambas as técnicas de codificação da forma de onda e de codificação da fonte. Com esta nova classe de codificadores conseguiu-se produzir boa qualidade da fala com taxas de bits intermediárias. A Figura 3.1 mostra o diagrama esquemático para a curva do comportamento típico da qualidade da fala versus a taxa de bits para as principais classes de codificadores da fala.

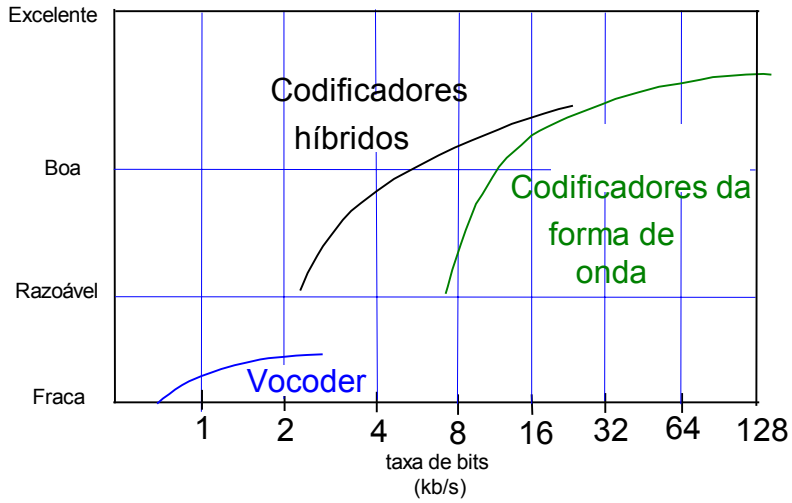

Figura 3.1 - Diagrama esquemático representando a qualidade da fala reproduzida em função da taxa de bits e do tipo de codificação, para sinais da fala na faixa telefônica. (Figura adaptada de [25]).

3.2 Codificadores de Forma de Onda

Codificadores de forma de onda tentam reproduzir amostra por amostra, o sinal original. A principal característica destes codificadores é a aproximação do sinal de saída ao de entrada à medida que o ruído de quantização diminui, o que permite uma verificação objetiva da reconstrução do sinal através da relação sinal/ruído. Eles são projetados para serem independentes do tipo de sinal e, portanto, podem ser utilizados para codificar uma ampla variedade de sinais. Geralmente são codificadores de complexidade baixa que produzem fala de alta qualidade em taxas acima de 16 kbits/s (ADPCM) até 64 kbits/s (PCM). A codificação da forma de onda pode ser realizada no domínio do tempo ou da freqüência.

3.2.1 Codificadores de forma de onda no domínio do tempo

Os codificadores no domínio do tempo executam o processo de codificação das amostras no tempo nos dados do sinal. Os métodos de codificação no domínio do tempo são [26]: Codificação por modulação de pulso "*Pulse Code Modulation*" (PCM), Codificação por modulação de pulso adaptativo (APCM), Codificação por modulação de pulso diferencial (DPCM), Codificação por modulação de pulso diferencial adaptativo (ADPCM), Codificação por modulação delta (DM), Codificação por modulação delta adaptativa (ADM) e Codificação Preditiva adaptativa (APC). Aqui são descritos alguns esquemas de codificação importantes no domínio do tempo.

Codificadores PCM – A codificação PCM é o tipo mais simples para a codificação da forma de onda. Essencialmente é apenas um processo de quantização amostra por amostra. Qualquer forma de quantização escalar pode ser usada neste esquema, mas a forma mais utilizada é a quantização logarítmica. A recomendação G.711 do CCITT[3.1] define a codificação PCM com 8 bits lei A e lei μ como método padrão de codificação para a fala em telefone.

Codificadores DPCM e ADPCM – Os codificadores PCM não fazem nenhuma suposição sobre a natureza da forma de onda a ser codificada, e trabalham bem para sinais diferentes dos sinais da fala. Entretanto, quando se trabalha com sinais da fala observa-se que existe uma grande correlação entre as amostras adjacentes. Esta correlação pode ser utilizada para reduzir a taxa de bits resultante. Um método simples de fazer isto é transmitir somente as diferenças entre as amostras adjacentes. O sinal diferença tem uma faixa dinâmica muito menor do que a do sinal da fala original e, portanto pode ser quantizada efetivamente usando um quantizador com uma quantidade menor de níveis de reconstrução. Neste método a amostra anterior está sendo usada para prever o valor da amostra atual. A predição pode ser melhorada se um bloco maior de amostras do sinal da fala for usado para fazer a predição. Esta técnica é conhecida como "*Differential Pulse Code Modulation"* (DPCM). Esta estrutura é mostrada na Figura 3.2.

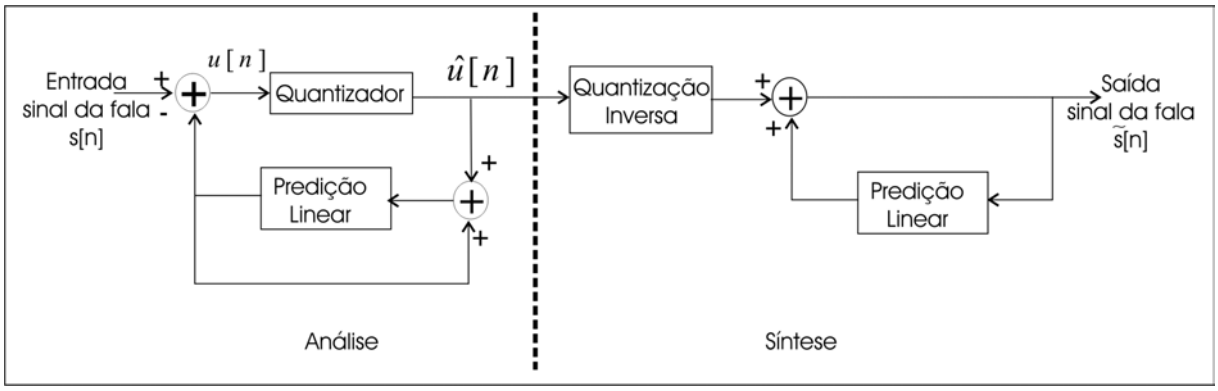

Figura 3.2 - Diagrama esquemático do sistema geral do DPCM: Codificador (análise) e decodificador (síntese).

Uma versão melhorada do DPCM é o "Adaptive DPCM" no qual o preditor e o quantizador são adaptados para as características localizadas do sinal de entrada. Existe um número de recomendações ITU baseadas em algoritmos ADPCM para a codificação da fala e áudio em banda estreita (taxa de amostragem de 8 kHz). Por exemplo, o G.726 operando em 40, 32, 24 e 16 kbits/s. A complexidade dos codificadores ADPCM é razoavelmente baixa.

[3.1] International Telegraph and Telephone Consultative Committee's (CCITT).

3.2.2 Codificadores de forma de onda no domínio da freqüência

Os codificadores de forma de onda no domínio da freqüência dividem o sinal em um número de componentes de freqüência separadas e as codifica separadamente. O número de bits usados para cada componente de freqüência pode variar dinamicamente. Codificadores no domínio da freqüência são divididos em dois grupos: codificadores sub-banda "*Sub-band Coding*" (SBC) e os codificadores por transformada adaptativa "Adaptive Transfom Coding" (ATC).

Codificação "*Sub-band Coding*" (SBC) – Na codificação SBC [2], os codificadores empregam poucos filtros passa faixa, isto é, um banco de filtros, para dividir o espectro do sinal de entrada em um número de faixas (tipicamente de 2 a 8), sinais sub-bandas, os quais são decimados e codificados separadamente, utilizando técnicas no domínio do tempo. No receptor, os sinais sub-bandas são decodificados, interpolados e somados para reconstruir o sinal de saída. A principal vantagem da codificação de sub-banda é que o ruído de quantização produzido em uma faixa é confinado àquela faixa. Como exemplo deste tipo de codificação, a recomendação G.722 do ITU-T para codificação de sinais de áudio de banda larga (7 kHz de largura de faixa amostrado em 16 kHz), a 64 kbits/s, divide o espectro em duas sub-bandas, codificando cada uma das bandas com codificadores ADPCM, derivados da recomendação G.727. O sinal da banda alta é codificado em 16 kbits/s e o sinal da banda baixa em 48 kbits/s, estando embutidos os codificadores de 40 kbits/s e 32 kbits/s para taxas menores.

Codificação "*Adaptive Transfom Coding*" (ATC) – Esta técnica envolve uma transformação em bloco de um segmento de uma janela de um sinal de entrada em faixas de freqüência, ou em outro domínio similar. Para codificar de modo eficiente, os bits disponíveis são distribuídos adaptativamente por faixa ou por coeficientes, de forma a codificar com maior acuidade as faixas (ou os coeficientes) mais importantes (com maior energia). No receptor o decodificador faz a transformação inversa para obter o sinal reconstruído. Várias transformadas como a Transformada de Fourier Discreta no Tempo "Discrete Fourier Transform" (DFT) ou a Transformada do Cosseno Discreta no Tempo "Discrete Cosine Transform" (DCT) podem ser usadas.

3.3 Codificadores de Fonte

Os codificadores de fonte ou *vocoders* usam um modelo de como a fonte foi gerada, e tentam extrair os parâmetros do modelo a partir do sinal a ser codificado. Estes parâmetros são

transmitidos ao decodificador. Assim os *vocoders* usam um modelo simplificado do mecanismo humano de produção da fala, mas não reproduzem fielmente o sinal de entrada, apenas o representam de modo a manter as suas características mais importantes, tais como a envolvente espectral, a estrutura fina do espectro e a energia global. O modelo é denominado de modelo do filtro-fonte de produção da fala como é mostrado na Figura 3.3. Este modelo supõe que a fala é produzida pela excitação de um filtro linear e variante no tempo (o trato vocal) com ruído branco para segmentos de sinais surdos, ou com um trem de impulsos para segmentos sonoros da fala. *Vocoders* operam em torno de 2 kbits/s e têm uma qualidade da fala sintética (não natural). Dependendo do método de extração dos parâmetros do modelo, vários tipos de vocoders tem sido desenvolvidos, como o vocoder de canal, vocoder homomórfico, vocoder de formantes e vocoder de predição linear (ou vocoder LPC).

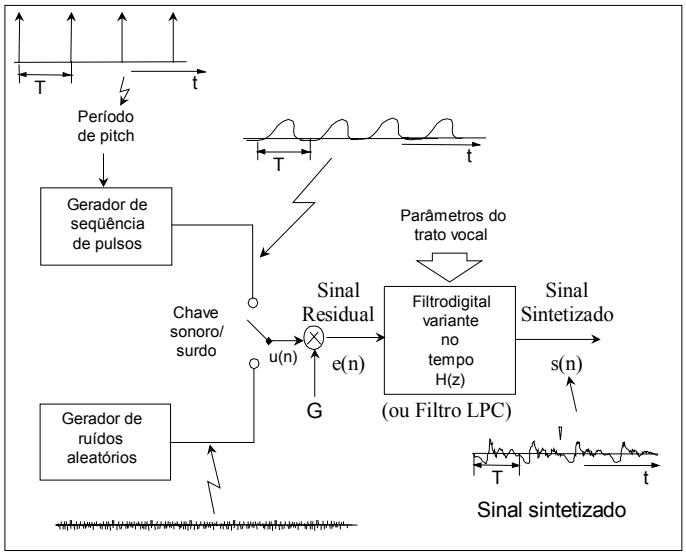

Figura 3.3 - Diagrama esquemático do modelo do filtro-fonte para a produção da fala em *vocoders*.

3.4 Codificadores Híbridos

Os codificadores híbridos surgiram na década de 80 com a pretensão de obter uma qualidade elevada com taxa de bits entre 3 e 16 kbits/s, faixa em que o sentimento generalizado era de que nenhum dos codificadores anteriores poderia funcionar com qualidade. Sua implementação em tempo real só é possível devido ao aparecimento dos processadores de sinal com grande capacidade de cálculo, considerando que utilizam algoritmos complexos. Estes codificadores aproveitam as vantagens dos dois tipos anteriores. Por um lado, são codificadores de forma de onda, porque tentam reproduzir o sinal original tanto em amplitude

como em fase, e por outro lado porque utilizam um modelo paramétrico com o objetivo de diminuir a taxa de bits.

Desta forma os codificadores híbridos tentam preencher a lacuna entre os codificadores de forma de onda e codificadores paramétricos. Codificadores de forma de onda são capazes de reproduzir o sinal da fala com uma boa qualidade com taxas de bits em torno de 16 kbits/s, por outro lado, os *vocoders* operando com taxas de bits muito baixas (2,4 kbits/s) não conseguem melhor que uma qualidade razoável, ou seja, uma qualidade não natural. De forma geral os codificadores híbridos podem ser esquematizados como mostra a Figura 3.4.

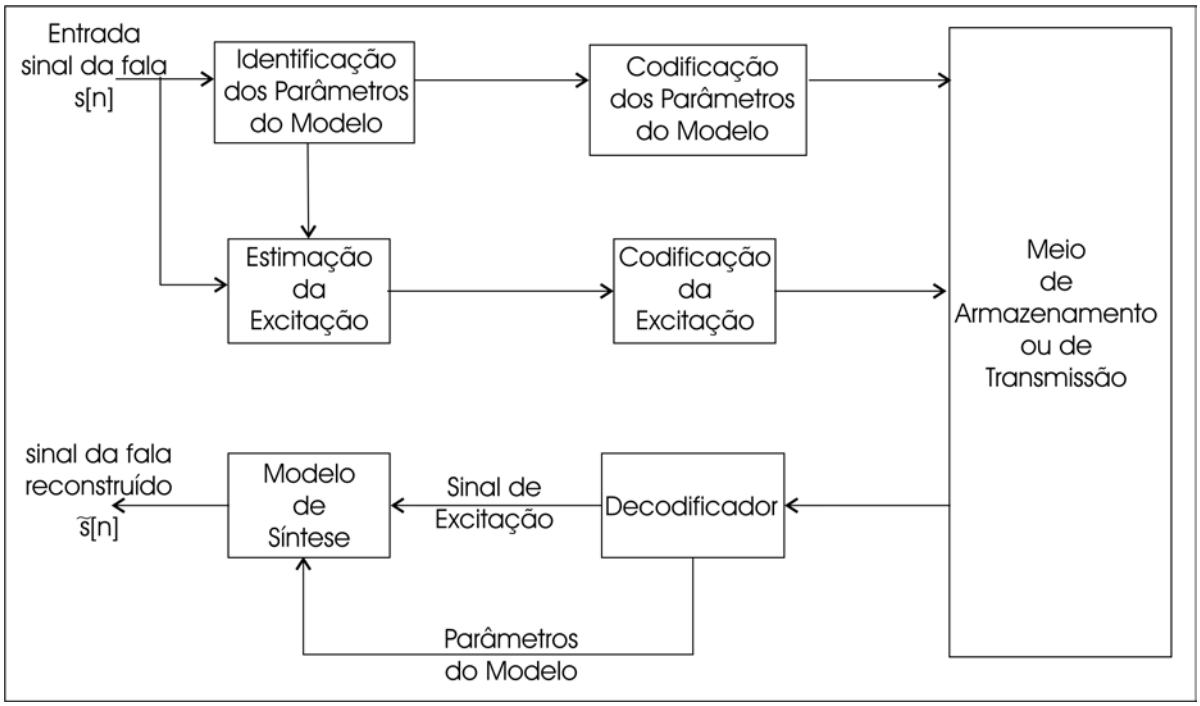

Figura 3.4 – Diagrama Esquemático para a codificação híbrida (codificador / decodificador).

Embora existam outras formas de codificadores híbridos, os mais utilizados e de maior sucesso são os *codificadores de análise por síntese* "*Analysis-by-Synthesis*" (AbS) no domínio do tempo. Tais codificadores usam o mesmo modelo de filtro do trato vocal encontrado nos *vocoders* LPC e por isso são também chamados de *codificadores de análise por síntese baseados em predição linear* (LPAS). Entretanto, ao invés de aplicar um modelo simples de dois estados, sonoro/surdo, para encontrar a entrada necessária para o filtro, o sinal de excitação é escolhido tentando aproximar a forma de onda da fala reconstruída tão próxima quanto possível do sinal original da fala. Um modelo geral para codificadores AbS é mostrado na Figura 3.5. Os codificadores LPAS foram introduzidos primeiramente por Atal e Remde [27] em 1982. Esse tipo de codificador tornou-se conhecido como codificador excitado por multi-pulso "*Multi-Pulse Excited Coder"* (MPE). Posteriormente foram introduzidos o codificador com excitação regular de impulsos "*Regular Pulse Excitated*" (RPE) [28, 29], e o

codificador por predição linear com excitação por código "*Code – Excited Linear Preditive Coders*" (CELP) [30, 31]. Muitas variações dos codificadores CELP tem sido padronizadas, incluído [26, 32] G.723.1 operando em 6.3/5.3 kbits/s, G729 operando em 8 kbits/s, G728 um codificador com baixo atraso operando em 16 kbits/s, e todos os padrões de codificação para telefonia móvel digital (celulares) [33, 34, 35] GSM, IS-54, IS-95, e IS-136.

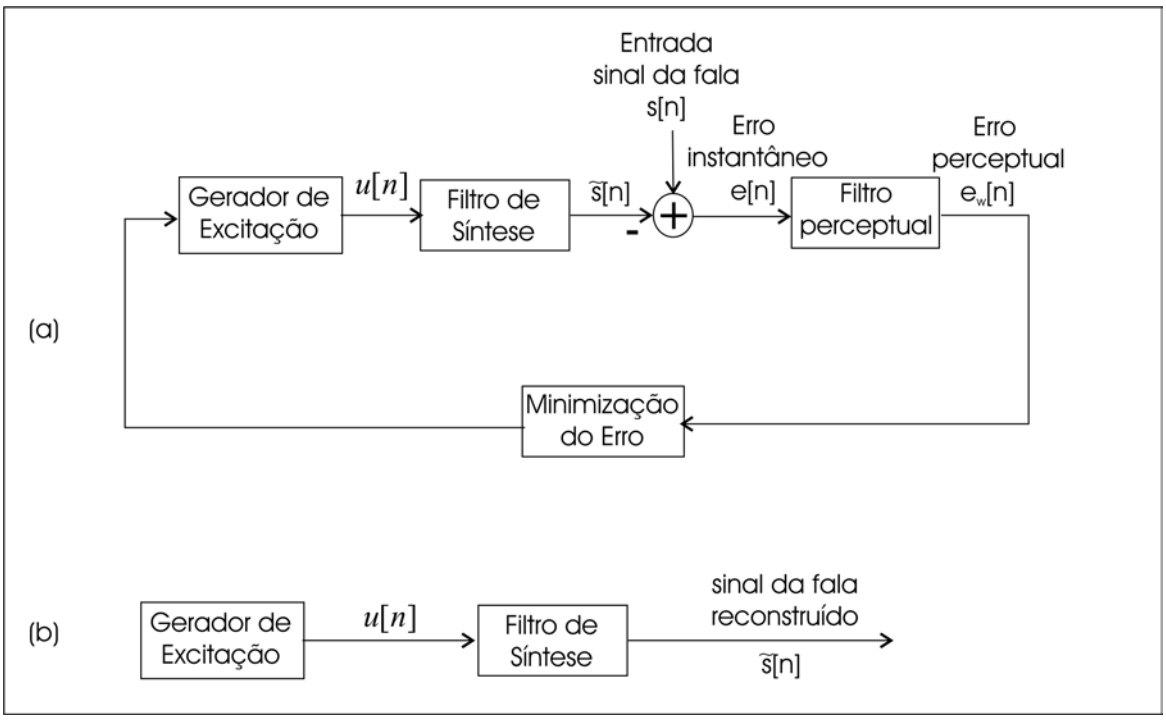

Figura 3.5 – Diagrama Esquemático para a codificação análise-por-síntese:
(a) codificador e (b) decodificador.

Embora os codificadores baseados em LPAS produzem uma alta qualidade da fala na faixa de 4-16 kbits/s, suas performances degradam rapidamente em torno de 4 kbits/s [36, 37] no ponto onde a performance da aproximação da forma de onda no domínio do tempo é deteriorada (mesmo com uma cuidadosa escolha para o critério do erro perceptual).

Uma alternativa viável aos codificadores LPAS, particularmente na faixa de 2-4 kb/s, abrange os codificadores que usam diretamente em sua análise as representações no domínio da freqüência com a incorporação de inovações para modelar de forma eficiente o sinal de excitação [38, 39]. Assim as técnicas utilizadas em codificação com baixas taxas de bits são: a codificação harmônica que inclui os codificadores sinusoidais e os codificadores "*Multi-Band Excitation"* (MBE); a codificação baseada em excitação mista "*Mixed Excitation Linear Prediction*" (MELP); e a codificação WI [24, 38, 39].

Codificadores Sinusoidais - Os codificadores sinusoidais são também codificadores híbridos, mas que trabalham no domínio da freqüência. A motivação para este modelo é que para segmentos sonoros a fala pode ser descrita pela sobreposição de sinusóides. Assumindo

periodicidade total, essas sinusóides são ainda harmonicamente relacionadas, originando a codificação harmônica. Em geral, os sinais sonoros da fala podem ser modelados como uma soma de ondas sinusoidais harmônicas espaçadas na freqüência fundamental, com fases amarradas à freqüência fundamental, enquanto que os sinais surdos da fala podem ser modelados como uma soma de sinusóides com fase aleatória [40].

Para a codificação "*Sinusoidal Transform Coding*" (STC) é suposto que ambos os segmentos da fala sonora e da fala surda podem ser representados por componentes de freqüências que tenham amplitudes e fases apropriadas, onde estas freqüências, amplitudes e fases são determinadas utilizando-se a transformada de Fourier de curto termo "*Short-Term Fourier Transform*" (STFT) de um quadro do sinal da fala. McAulay e Quatieri [40] sugeriram então selecionar os picos no espectro da STFT com o objetivo de determinar as freqüências componentes da forma de onda da fala. Essas freqüências têm amplitudes e fases associadas que podem ser combinadas para reproduzir a forma de onda da fala com o número de picos determinados na STFT. A redução da taxa de bits é atingida forçando o modelo sinusoidal ter fase zero [41].

Codificadores "*Multi-Band Excitation*" (MBE) - Os codificadores MBE [42, 43, 44, 45] são codificadores no domínio da freqüência que incorporam uma inovação para melhorar o modelo de excitação. O modelo de excitação é misto, permitindo que ambas componentes, harmônicas e aleatórias, participem de um mesmo quadro da fala. Para sinais sonoros da fala, uma seqüência periódica de impulsos de excitação corresponde no domínio da freqüência a uma seqüência periódica de impulsos no domínio da freqüência, espaçadas em harmônicas do *pitch*. O modelo MBE divide o espectro e as sub-bandas em múltiplos da freqüência fundamental (ou freqüência de pitch). A maneira na qual o vocoder MBE representa as informações de freqüência do trato vocal pode ser pensada como um vocoder de canal que tem todos os canais centrados em harmônicas da freqüência fundamental (ou freqüência de pitch). O modelo MBE permite separar a decisão sonoro/surdo para cada canal (ou grupo de canais) em cada quadro do sinal da fala. Isto permite uma representação do sinal de excitação mais fiel do que a decisão simples dos *vocoders*.

Um codificador "*Improved Multi-Band Excitation coder*" (IMBE) [46] operando a uma taxa de 4,15 kbits/s (MOS~3.3) foi selecionado pelo Inmarsat como padrão para a comunicação de voz por satélite.

Codificadores "Mixed Excitation Linear Prediction -" MELP – A codificação MELP é um modelo que reduziu a taxa de bits abaixo das taxas obtidas com a codificação CELP enquanto melhorou a qualidade relativa ao dois estados da codificação LP [47]. MELP emprega a análise LPC para modelar o espectro de curto termo, mas evita a simples decisão de sonoro/surdo para um quadro completo do sinal da fala. Ele modela a excitação, em faixas

separadas através do espectro de freqüência, como uma combinação de componentes periódicas e ruidosas, com sua relativa contribuição baseada nas informações de sonoridade (no grau de intensidade da sonoridade). Esta maneira de abordar modela com mais eficiência os segmentos da fala que têm excitação mista como o fonema /z/ que é classificado como um fricativo sonoro, por exemplo, em transições entre sonoro e surdo. O codificador é similar ao codificador básico LP de dois estados com a adição das características de excitação mista, pulsos aperiódicos, e filtro de dispersão de pulso. Em 1997, o DoD[3.2] normalizou o codificador MELP produzindo uma taxa de bits de 2,4 kbits/s [48, 49] para substituir as normas FS-1015 e FS-1016.

Codificadores WI – A codificação da fala por interpolação de ondas é uma tentativa para combinar aspectos de análise no domínio do tempo e no domínio da freqüência. A técnica "*Prototype Waveform Interpolation*" (PWI) [50, 51] e a técnica "*Time Frequency Interpolation*" (TFI) [51, 52] foram as técnicas precursoras da técnica WI [15, 54]. Um codificador WI implementado em 2,4 kbits/s demonstrou alta qualidade do sinal sintetizado da fala com MOS ~ 3.5 [55, 56]. Nesses últimos anos muitas técnicas têm sido sugeridas para o usar o esquema WI [57, 58, 59, 60, 61]. Embora a interpolação da forma de onda seja executada usualmente no domínio da freqüência, a interpolação no domínio do tempo tem sido implementada com bons resultados [62]. Similaridades e diferenças entre WI e STC são também examinadas em [63] e [15]. O codificador WI que é objeto de estudo desta tese é abordado em detalhes no capítulo 4.

3.5 Considerações Finais do Capítulo

Neste capítulo foi apresentada uma visão geral das técnicas de codificação da fala que são utilizadas atualmente, e sobre aquelas que poderão ser utilizadas no futuro. Foi apresentada uma classificação geral onde, os codificadores da fala são divididos tradicionalmente em duas classes – *codificadores de forma de onda* e *codificadores de fonte* (que também são conhecidos como codificadores paramétricos, os vocoders "*voice coders*"). Os *codificadores de forma de onda* produzem uma fala de alta qualidade, mas a custa de taxas de bits elevadas e os *vocoders* são usados em taxas de bits muito baixas, mas tendem a produzir uma fala de qualidade sintética. Mais recentemente foi introduzido uma nova classe de codificadores, chamados de codificadores híbridos, os quais usam ambas as técnicas de codificação da forma de onda e de codificação da fonte. Com esta nova classe de codificadores conseguiu-se produzir boa qualidade da fala em taxas de bits intermediárias.

[3.2] "Department of defense (DoD - USA)"

O objetivo principal foi mostrar a seqüência do desenvolvimento e o aparecimento das técnicas de codificação da fala, apontando alguns codificadores padronizados nos sistemas de comunicação e situando o contexto do início e o desenvolvimento da técnica de codificação da fala por interpolação de ondas, a técnica WI. No próximo capítulo é apresentada a técnica WI com o objetivo de permitir a visualização de um algoritmo para a implementação de um simulador para o codificador/decodificador WI básico, como parte deste trabalho.

Capítulo 4

CODIFICAÇÃO DA FALA POR INTERPOLAÇÃO DE ONDAS
"Waveform Interpolation Speech Coding" (WI)
A Técnica e o Codificador WI

4.1 Introdução

Este capítulo apresenta a descrição da técnica de codificação da fala por interpolação de ondas, denominada de técnica WI, e a descrição dos processos relativos ao codificador/decodificador básico e convencional, utilizando a técnica WI. A denominação de codificador básico indica que o codificador estará operando em uma fase (ou camada) onde os parâmetros são transmitidos sem compressão.

Este trabalho, a princípio, está focalizado no estágio de análise (durante a codificação), com objetivo de extrair os parâmetros do sinal, e no estágio de síntese do sinal da fala (durante a decodificação) para a reconstrução do sinal. Portanto, não é necessário trabalhar com a compressão (ou codificação) dos parâmetros.

Neste capítulo são apresentadas, a descrição geral da técnica WI e a descrição do codificador WI que compreende a estrutura básica, a representação da forma de onda característica e os estágios de análise e de síntese.

Na descrição do codificador/decodificador procurou-se partir de uma visão geral, focalizando as estruturas básicas, para uma visão em detalhes, focalizando em cada processo os parâmetros de entrada e saída, os procedimentos envolvidos, e apresentando alguns algoritmos que não estão disponíveis na literatura, que facilitam a compreensão e permitem executar a implementação de um simulador para o codificador WI básico e convencional, ou seja, um sistema WI de análise-sintese para o sinal da fala.

Para a descrição da técnica foram usados como fonte principal de pesquisa os trabalhos apresentados por W. B. Kleijn em [15, 50, 51, 54, 55, 56, 57] e para auxiliar na implementação do simulador o trabalho apresentado por E. L. T. Choy, em [14].

4.2 A Técnica da Codificação Da Fala Por Interpolação De Ondas
"Waveform Interpolation Speech Coding" - A Técnica WI

Durante segmentos sonoros, o sinal da fala é quase periódico. Começando em um instante de tempo qualquer, é fácil identificar o primeiro ciclo do pitch, o segundo ciclo, e assim por

diante. Quando se compara esta seqüência de formas de onda de ciclos do pitch, observa-se que o formato geral evolui lentamente em função do tempo, ou seja, existe uma grande correlação entre elas. A evolução lenta da forma de onda sugere que, no lado do codificador, a forma de onda do ciclo do pitch pode ser extraída em intervalos regulares no tempo, tipicamente a cada 20 ms, transmitida e então, no lado do decodificador, o sinal original pode ser obtido com uma boa aproximação, por meio de interpolação das formas de ondas intermediárias entre as formas de onda transmitidas, como mostrado na Figura 4.1. Este raciocínio foi a motivação original para o desenvolvimento da técnica de codificação da fala por interpolação de ondas "*waveform interpolation (WI) speech coding*" que são descritos a seguir [15].

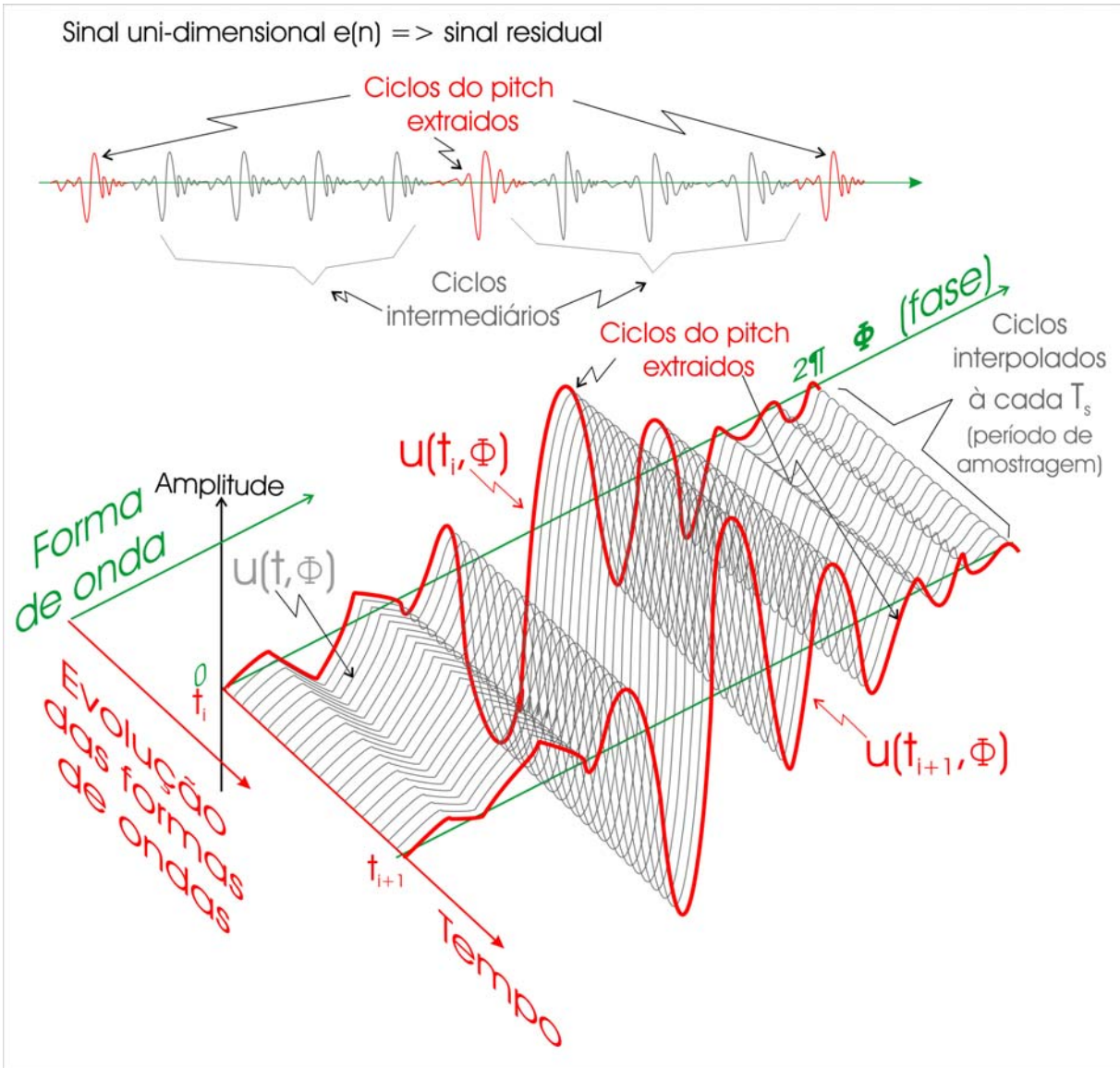

Figura 4.1 - Motivação para a codificação WI.

"Para sinais sonoros da fala, a forma de onda do ciclo do pitch descreve as características essenciais do sinal da fala. Se a forma de onda do ciclo do pitch e a fase correspondente são conhecidas em cada instante do tempo, então é possível reconstruir o sinal da fala sem distorções, como mostrado na Figura 4.2. Os sinais sonoros da fala podem, portanto ser representados como uma superfície bi-dimensional, $u(t,\phi)$, onde a forma de onda é disposta ao longo do eixo da fase (ϕ) e a evolução da forma de onda ao longo do eixo do tempo (t), com os eixos (ϕ) e (t) ortogonais entre si. Enquanto esta descrição é mais natural para a fala sonora, ela também é válida para sinais não periódicos como a fala surda. Por esta razão a forma de onda disposta ao longo do eixo ϕ é referida como *forma de onda característica*, ("*characteristic waveform*" (CW)). Na maioria das implementações WI, a forma de onda característica é normalizada em duração e periódica em 2π. Durante a fala sonora a forma de onda característica apresenta a forma de um ciclo do pitch. Junto com o período de pitch, que também varia em função do tempo, a *forma de onda característica evolutiva* $u(t,\phi)$ pode ser utilizada para descrever o sinal da fala" [15].

"A forma de onda característica, CW, evolui lentamente, ou suavemente, para sinais sonoros da fala e rapidamente, ou abruptamente, durante sinais surdos da fala. De uma forma geral, as componentes quase-periódicas do sinal da fala correspondem à *componente* da forma de onda característica *evolutiva lenta* (ou componente de evolução lenta), enquanto as componentes não periódicas (ou ruidosas) do sinal da fala correspondem à *componente evolutiva rápida* (ou componente de evolução rápida). Assim, a forma de onda característica evolutiva não apresenta periodicidade, mas ao invés disso, a taxa de mudança da forma de onda característica é função do nível de periodicidade do sinal da fala original em uma dimensão" [15].

"Dois tipos de espectro de Fourier estão associados com a *forma de onda característica evolutiva*. Uma transformada de Fourier de $u(t,\phi)$ ao longo do eixo ϕ (ao longo do qual a forma de onda é disposta), em um determinado tempo t mostra um espectro de curto termo associado com aquele instante de tempo. Este espectro é semelhante à transformada de Fourier de curto termo do próprio sinal da fala. O comprimento da forma de onda característica geralmente é normalizado, resultando em uma escala de freqüência diferente" [15].

"Em contraste, a transformada de Fourier ao longo do eixo dos tempos t para um determinado valor de ϕ mostra um espectro da evolução das freqüências, o qual está relacionado com a taxa de evolução da forma de onda característica. Para segmentos sonoros da fala, a maior parte da energia em tal espectro de evolução está abaixo de 20 Hz. A largura de faixa do espectro de evolução é função do período de pitch. Para uma reconstrução perfeita, o sinal deveria ser amostrado uma vez por período de pitch, denotado aqui por $p(t)$, o que implica que a largura de faixa é no máximo $1/2p(t)$" [15].

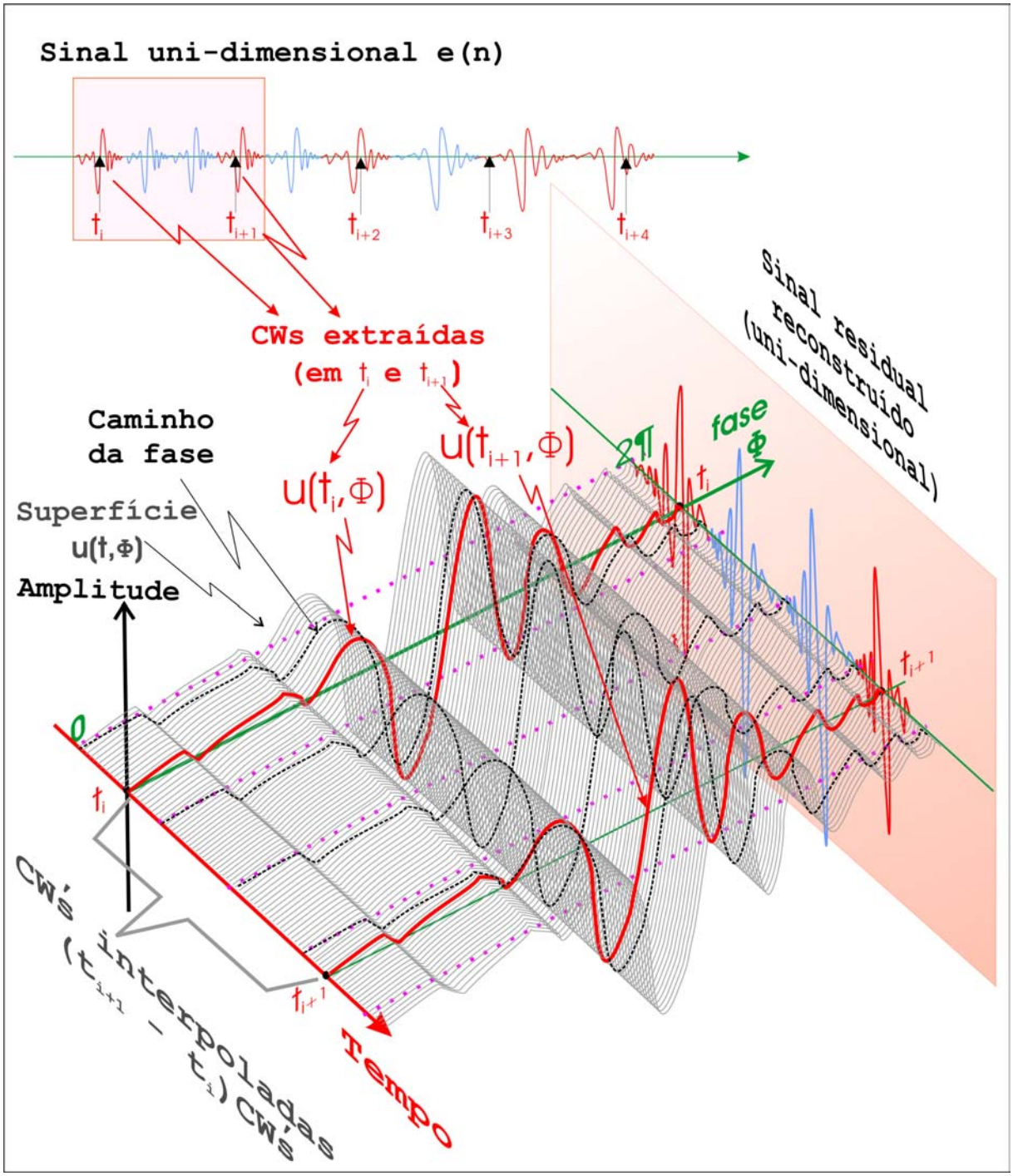

Figura 4.2 - Sistema básico da técnica de codificação WI da fala: extração, interpolação e reconstrução do sinal.

"A natureza diferente na percepção das componentes sonoras e surdas para sinais da fala provê uma forte motivação para a separação destas componentes na codificação da fala com taxa de bits baixa. As formas de ondas características evolutivas provêem um domínio ideal para se fazer a separação das componentes usando filtragem linear simples. A filtragem passa baixa da função $u(t,\phi)$ ao longo do eixo dos tempos t resulta na *forma de onda evolutiva lenta* "*slowly evolving waveform*" (SEW), que corresponde à componente quase-periódica do

sinal da fala. A filtragem passa alta da função $u(t,\phi)$ ao longo do eixo dos tempos t resulta na *forma de onda evolutiva rápida* "*Rapidly evolving waveform*" (REW), que corresponde à componente ruidosa do sinal da fala" [15].

"Os princípios descritos anteriormente foram desenvolvidos visando principalmente o aumento da eficiência da codificação da fala. É natural o uso do protótipo, a forma de onda do ciclo de pitch, para descrever a fala sonora, reduzir a taxa de amostragem (sub amostrar) e interpolar estas formas de ondas para reduzir a taxa de bits. Basicamente, a forma de onda era extraída (amostrada) e transmitida a uma taxa baixa (tipicamente de 50 Hz), e a função $u(t,\phi)$ era então reconstruída no receptor por meio de interpolação linear nos codificadores de forma de onda operados com taxas entre 2,4 e 4 kb/s" [15]. Isto foi aplicado ao codificador desenvolvido na primeira versão da técnica de codificação da fala por interpolação de ondas, introduzida por W. B. Kleijn em 1991, e denominada de "*Prototype Waveform Interpolation*" (PWI). Ela foi aplicada para codificar somente segmentos sonoros da fala. O codificador PWI foi usado em combinação com outra técnica como CELP para a codificação dos segmentos surdos.

"Codificadores WI que amostram as formas de ondas características com baixas taxas, os PWI's, fazem a suposição de que o sinal bidimensional $u(t,\phi)$ evolui lentamente na direção de t, isto é, que $u(t,\phi)$ tem uma largura de faixa de evolução estreita. Por esta razão, as implementações WI antes de 1994, ou seja, os codificadores PWI, foram condicionados para a fala sonora. Uma vez que a fala sonora geralmente contém componente de ruído, percebeu-se nesta época que a qualidade do sinal reconstruído aumentava se a taxa de amostragem das formas de ondas características era aumentada [53, 57]. Entretanto, sem promover mudanças no algoritmo, o aumento da taxa de amostragem está associado com um aumento proporcional na taxa de bits. Um método para contornar o problema foi então aplicar a transformada de Fourier bi-dimensional em blocos de $u(t,\phi)$ e combinar com uma alocação não uniforme de bits no domínio da transformada resultante. O método de codificação WI usando estas transformadas bidimensionais assemelha-se de algum modo ao codificador de transformada com pitch sincronizado [64]" [15].

"Assim, a partir de 1994, a técnica PWI foi aperfeiçoada para codificar ambos os sinais da fala sonora e surda com bons resultados, quando passou a ser denominada de WI "*Waveform Interpolation (WI) speech Coding*" [54, 58]. Através da separação das componentes quase-periódicas e das componentes ruidosas antes da quantização, os benefícios do aumento na taxa de amostragem podem ser obtidos sem um significante acréscimo na taxa de bits. Primeiro, o sinal $u(t,\phi)$ é amostrado com uma alta taxa (tipicamente de 400 Hz), e então é separado em SEW e REW. A SEW pode ser sub amostrada (tipicamente a uma taxa de 50 Hz), e codificada com métodos desenvolvidos anteriormente para sistemas operando na fala sonora com taxa de bits baixa. Para REW, as propriedades do sistema auditivo humano podem ser exploradas. Para sinais surdos da fala, somente o envelope da

potência do sinal, com uma resolução de aproximadamente 5 ms, e uma descrição do espectro de potência com baixa resolução são perceptivamente significantes [65]**.** Na codificação WI esta noção se generaliza para todas as componentes ruidosas do sinal da fala, o que permite a codificação com taxa de bits baixa para a REW" [15].

4.3 Codificador por Interpolação de Formas de Ondas - O codificador WI

Nesta seção é apresentada a estrutura básica de um codificador da fala por interpolação de ondas, codificador WI, com as principais características para uma implementação típica. Será descrito um codificador WI operando sem o uso da codificação (quantização) dos parâmetros, pois este trabalho, a princípio, não necessita de trabalhar com sinais quantizados.

O objetivo é determinar um algoritmo para a implementação da estrutura básica de um sistema de codificação WI convencional, análise-síntese, sem o uso da codificação paramétrica.

São descritos os processos e os procedimentos envolvidos, e alguns algoritmos que não estão disponíveis na literatura.

Na seção 4.3.1 é apresentada a descrição da estrutura básica de um codificador da fala usando a técnica de codificação por interpolação de ondas.

Na seção 4.3.2 é apresentada a descrição da representação das formas de ondas, as "CW's", ou seja, o modelo matemático para as "CW's", onde é utilizada a representação através da Série Discreta de Fourier no Tempo, ou "*Discrete-Time Fourier Series*" (DTFS).

Nas seções 4.3.3 e 4.3.4 são apresentados os estudos, a descrição da técnica WI e dos processos, alguns algoritmos e procedimentos relacionados com o processamento dos sinais para o estágio de análise e o estágio de síntese do codificador/decodificador WI.

A intenção é que a partir destas seções seja possível compreender a técnica WI e permitir a visualização de um algoritmo para um codificador WI básico (sem codificação de parâmetros) que possa ser implementado.

4.3.1 Estrutura Básica

Um sistema de codificação por interpolação de ondas, como qualquer outro sistema de codificação da fala consiste de um codificador e um decodificador, como mostrado na Figura. 4.3. O codificador recebe o sinal digital original da fala em sua entrada e produz uma seqüência de bits "bitstream" em sua saída a uma determinada taxa de transmissão (ou taxa de bits). A seqüência de bits é transmitida ao decodificador que a decodifica e reconstrói o sinal da fala.

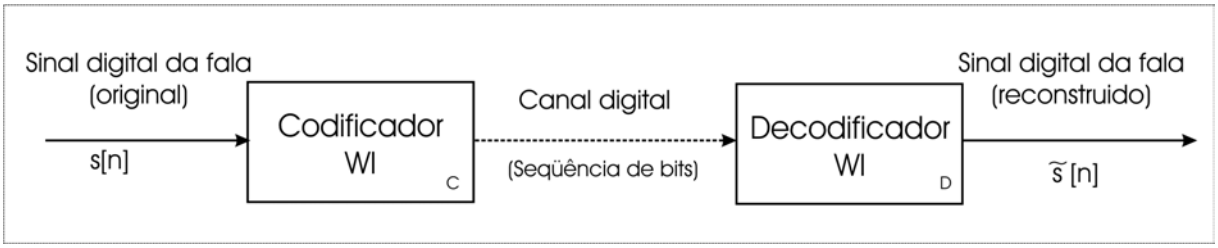

Figura 4.3 - Sistema de codificação da fala.

A Figura 4.4 apresenta um diagrama[4.1] esquemático geral do sistema de codificação WI em alto nível de abstração. O sistema WI pode ser dividido em duas camadas distintas: a *camada interna* e a *camada externa* que se estendem, cada uma delas, desde o codificador até ao decodificador. A *camada externa*, também denominada de *camada de análise-síntese,* compreende o processo C100, bloco de análise e o processo D200, bloco de síntese, localizados respectivamente, no codificador e decodificador. A *camada interna*, também denominada de *camada de codificação (ou de quantização),* compreende o processo C300, bloco de codificação de parâmetros (ou quantizador) e o processo D400, bloco de decodificação de parâmetros, também localizados respectivamente, no codificador e decodificador WI.

De uma forma geral no sistema codificador/decodificador WI são executados os seguintes processos:

1) Processo C100 (bloco de análise): Recebe o sinal digital da fala original, faz a análise LPC e obtém o sinal residual correspondente. Faz a estimação do período de pitch e o sinal residual é então decomposto em uma série de CW's. As CW's são alinhadas e normalizadas em potência para representar com precisão a superfície bidimensional $u(t,\phi)$ que mostra a evolução das formas de onda característica, as CW's.
2) Processo D200 (bloco de síntese): Faz o inverso do bloco de análise. O sinal residual é reconstruído a partir das CW's e enviado a um filtro de síntese onde o sinal da fala é finalmente reconstruído.
3) Processo C300 (bloco de codificação de parâmetros): Recebe os parâmetros WI, processa a decomposição SEW/REW e realiza a codificação (quantização) dos parâmetros.

[4.1] Para clareza e continuidade, será adotada uma notação semelhante à utilizada em [14]. Cada bloco funcional (que é referido como **processo** no resto do texto) no diagrama esquemático do sistema WI é identificado por quatro dígitos. O primeiro é uma letra para indicar se o bloco pertence ao codificador, C, ou ao decodificador, D. Cada um dos três dígitos na seqüência é um número que indica um nível de embutimento (ou encapsulamento). Por exemplo, um processo rotulado por **C172** indica que o processo está embutido no **processo C170** que está embutido no **processo C100** embutido no **Codificador WI, C.**

4) Processo D400 (bloco de decodificação de parâmetros): Recebe os parâmetros quantizados, realiza a decodificação e reconstrói as CW's a partir das componentes SEW/REW.

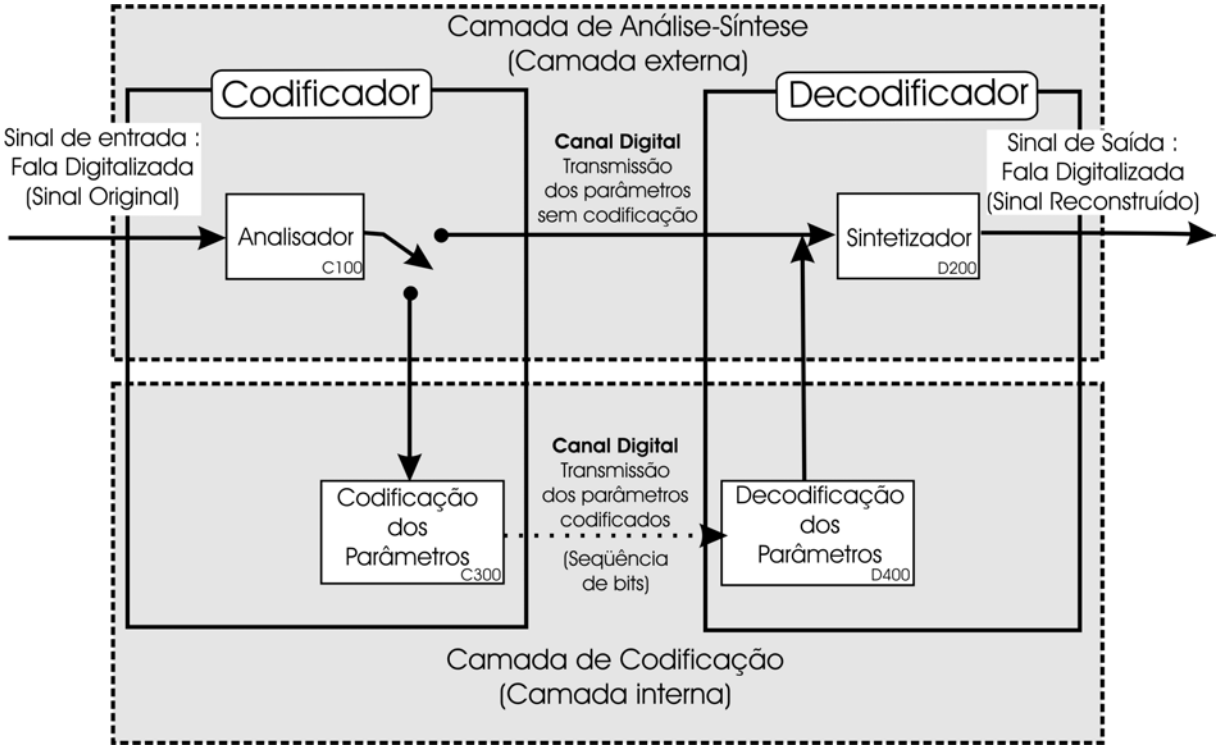

Figura 4.4 - Diagrama esquemático geral do sistema de codificação WI [14].

O sistema apresenta os caminhos I ou II, como opções para o processamento do sinal da fala, como mostrado na Figura 4.5. No caminho I, de uso normal para um codificador/decodificador, o sinal é recebido pelo codificador no bloco de análise, processo C100, que determina os valores dos parâmetros WI. Em seguida os parâmetros são codificados (ou quantizados) através do processo C300 e dispostos em uma *seqüência de bits* com um número de bits pré-determinado. A *seqüência de bits* resultante é transmitida para o decodificador por um canal digital ou por um meio de armazenamento. A *seqüência de bits* com os parâmetros codificados (ou quantizados) é então recebida pelo decodificador WI por meio do processo D400, onde são decodificados, e enviados para o processo D200, onde são então convertidos em um sinal da fala no bloco de síntese.

No caminho II, o processamento é realizado somente na camada de análise-síntese. Desta forma os parâmetros WI obtidos no bloco de análise, processo C100, são transmitidos sem codificação, diretamente para o bloco de síntese, processo D200, onde são convertidos em um sinal da fala.

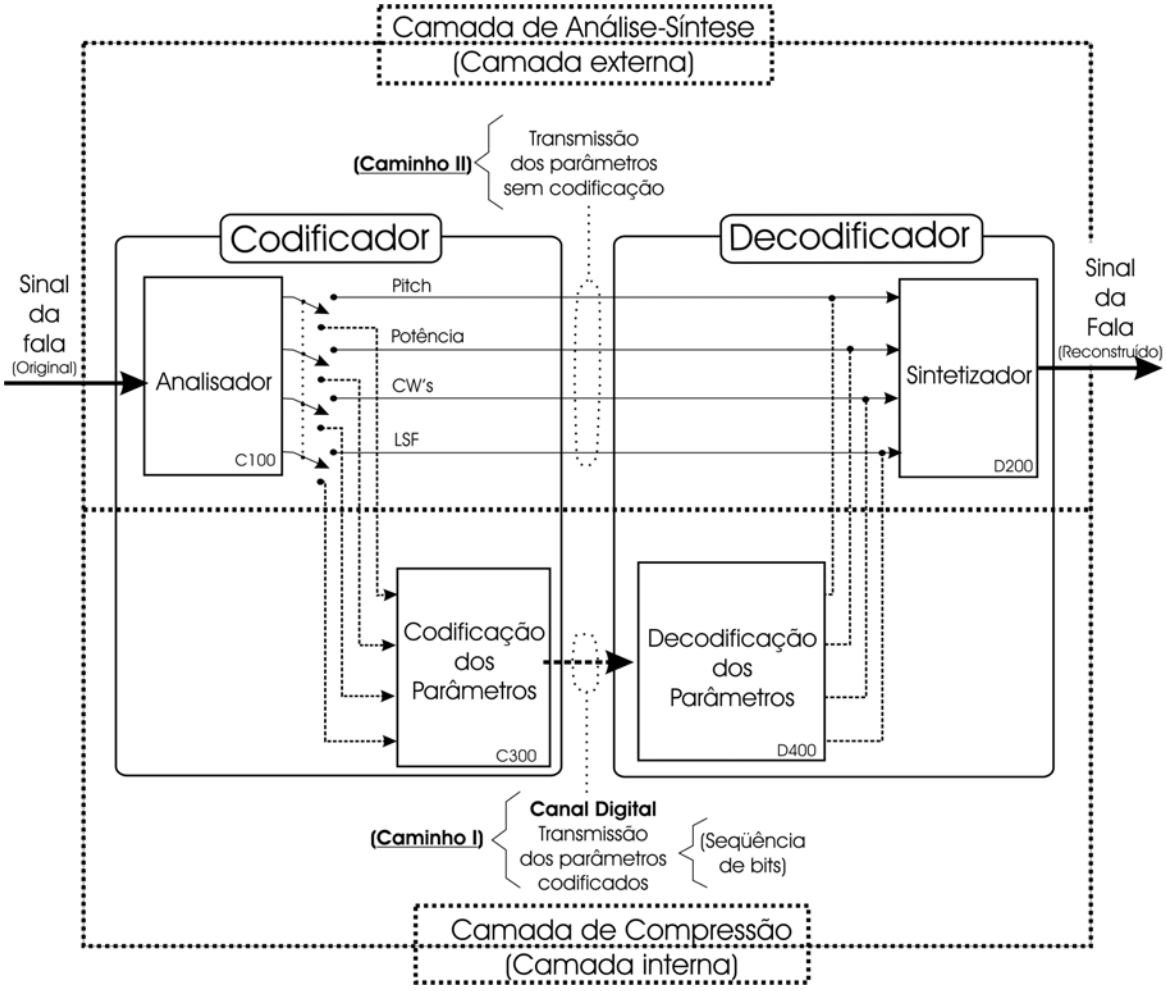

Figura 4.5 - Diagrama esquemático geral do sistema de codificação WI. Os parâmetros do sistema WI.

Nas seções que seguem os objetivos principais são o estudo e a descrição da técnica e dos processos relacionados com o processamento dos sinais da fala na camada de análise-síntese (caminho II), visando a visualização de um algoritmo para a implementação do codificador/decodificador WI básico (camada análise-síntese) e convencional (sem reconstrução perfeita).

4.3.2 Representação da forma de onda característica

A forma de onda característica (CW) é uma seqüência real, discreta e periódica no tempo com um período de onda igual a um período de pitch. Normalmente a CW é um segmento extraído do sinal da fala ou do sinal residual. Ela pode ser representada, em uma dimensão, por $v[m]$ com período de pitch P, como mostrado na Figura 4.6. Assim

$$v[m] \in \Re, \quad \text{onde} \quad m=0,1,...,P-1. \tag{4.1}$$

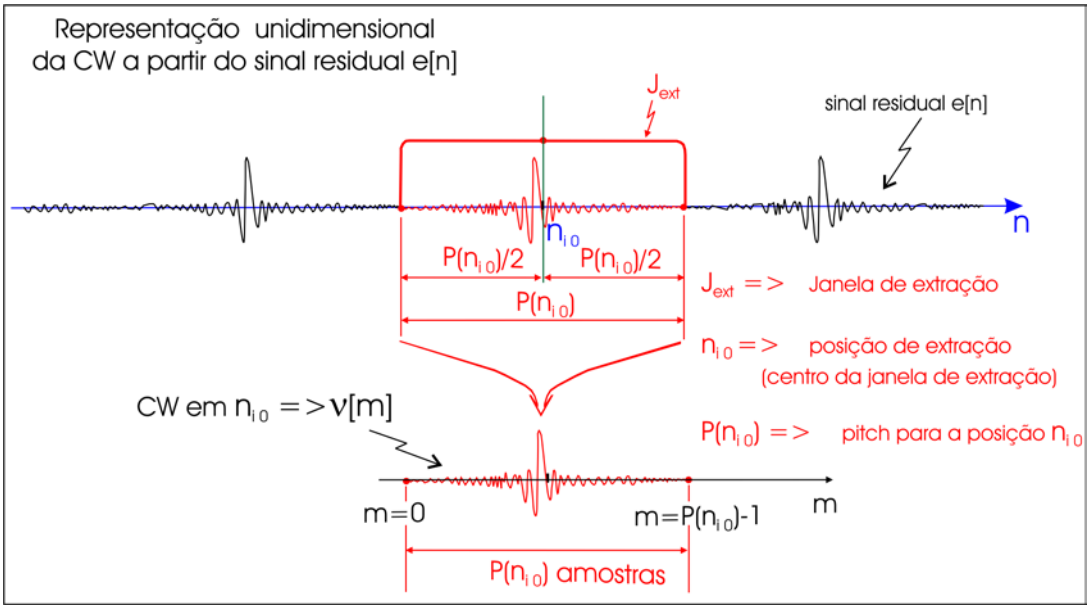

Figura 4.6 - Diagrama esquemático – Representação da forma de onda característica, CW, em uma dimensão.

No domínio da freqüência a CW pode ser representada pela Série Discreta de Fourier no Tempo, "Discrete-Time Fourier Series – DTFS",

$$v[m] = \sum_{k=0}^{\lfloor P/2 \rfloor} [A_k \cos(\frac{2\pi km}{P}) + B_k \operatorname{sen}(\frac{2\pi km}{P})], \qquad 0 \le m < P,$$
$$\begin{cases} com \quad \lfloor P/2 \rfloor = P/2, \quad para \ P \ par, \quad e \\ \lfloor P/2 \rfloor = (P-1)/2, \quad para \ P \ ímpar. \end{cases} \tag{4.2}$$

onde $\{A_k\}$ e $\{B_k\}$ são os coeficientes da DTFS que podem ser calculados usando o seguinte conjunto de equações:

$$A_k = \frac{\chi_k}{P} \sum_{m=0}^{P-1} \left[v[m] \cos\left(\frac{2\pi km}{P}\right) \right] \quad onde \quad 0 \le k \le \lfloor P/2 \rfloor$$
$$B_k = \frac{\chi_k}{P} \sum_{m=0}^{P-1} \left[v[m] \operatorname{sen}\left(\frac{2\pi km}{P}\right) \right] \quad e \quad \chi_k = \begin{cases} 2, & para \quad 0 < k < P/2 \\ 1, & para \quad k = 0 \ e \ k = P/2. \end{cases} \tag{4.3}$$

Assim uma CW, $v[m]$, no domínio do tempo pode então ser representada no domínio da freqüência pelos coeficientes de Fourier $\{A_k, B_k\}$.

Para uma seqüência de CW's a representação se torna bidimensional, como mostrado na Figura 4.7. Esta representação é obtida adicionando o tempo discreto n_i a todos os parâmetros da Equação (4.2) que podem variar com o tempo. Os parâmetros são

$$\{A_k\} \Rightarrow \{A_k(n_i)\},\ \{B_k\} \Rightarrow \{B_k(n_i)\} \quad \text{e} \quad P \Rightarrow P(n_i), \text{ onde } i = i_o + jb,\ j \in Z \text{ e}$$

b é a distância, em número de amostras, medida entre dois pontos de extração das CW's. A Equação (4.2) pode ser reescrita como,

$$v[n_i, m] = \sum_{k=1}^{\lfloor P(n_i)/2 \rfloor} [A_k(n_i)\cos(\frac{2\pi km}{P(n_i)}) + B_k(n_i)\,\text{sen}(\frac{2\pi km}{P(n_i)})], \quad 0 \le m < P(n_i), \tag{4.4}$$

$$\text{onde} \quad \begin{cases} \lfloor P(n_i)/2 \rfloor = P(n_i)/2, & \text{para } P \text{ par, e} \\ \lfloor P(n_i)/2 \rfloor = (P(n_i)-1)/2, & \text{para } P \text{ ímpar.} \end{cases}$$

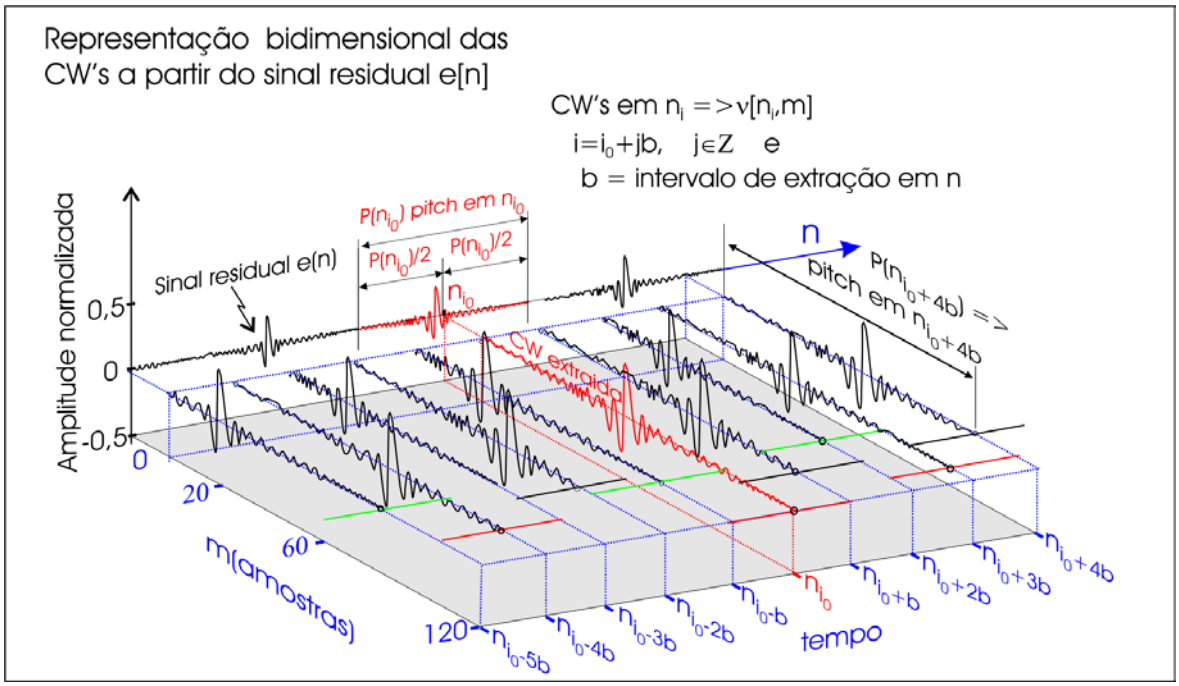

Figura 4.7 - Diagrama esquemático – Representação das formas de ondas características, CW's, em duas dimensões => $v[n_i, m]$.

Os coeficientes A_0 e B_0 são ignorados na Equação (4.4). B_0 na Equação (4.3) é redundante, pois sen(0)=0. Já A_0 representa a componente CC do sinal e não tem importância perceptível relevante [14]. Assim o índice k começa em 1 e não em zero.

A Equação (4.4) é uma representação do sinal bidimensional onde as variáveis independentes são m e n_i.

As CW's são dispostas individualmente ao longo do eixo dos tempos m (*eixo de fase*) e a forma de onda, ou formato da CW evolui no tempo ao longo do eixo n_i (*tempo de evolução das formas de ondas*, CW's). Assim m é o índice discreto que indica a posição temporal das amostras de voz ou do sinal residual dentro de um ciclo, ou dentro de uma CW, enquanto n_i também é um índice discreto, associado à cada CW, que indica a posição temporal de extração da CW (posição da amostra localizada na porção média da CW) em relação às amostras do sinal de voz ou residual.

O comprimento das CW's na Equação (4.4) é variante no tempo, pois é função do pitch $P(n_i)$ que também varia de CW para CW. Na técnica WI é conveniente que as CW's tenham o mesmo comprimento. Assim, a normalização é feita pela substituição de

$$\left\{\begin{array}{l} 0 \le m < P(n_i) \\ \\ 0 \le \phi < 2\pi \end{array}\right\} \Rightarrow \quad \phi = \phi(m) = \frac{2\pi m}{P(n_i)} \tag{4.5}$$

na Equação (4.4), que se torna:

$$v[n_i, \phi] = \sum_{k=1}^{\lfloor P(n_i)/2 \rfloor} [A_k(n_i)\cos(k\phi) + B_k(n_i)\,\mathrm{sen}(k\phi)], \quad 0 \le \phi(.) < 2\pi,$$

$$\textit{onde} \quad \left\{\begin{array}{l} \lfloor P(n_i)/2 \rfloor = P(n_i)/2, \quad \textit{para } P \textit{ par}, \quad e \\ \\ \lfloor P(n_i)/2 \rfloor = (P(n_i)-1)/2, \quad \textit{para } P \textit{ ímpar} \end{array}\right. \tag{4.6}$$

onde $\phi = \phi(m)$ é a fase. Assim todas as CW's ficam com o mesmo comprimento 2π, como mostrado na Figura 4.8.

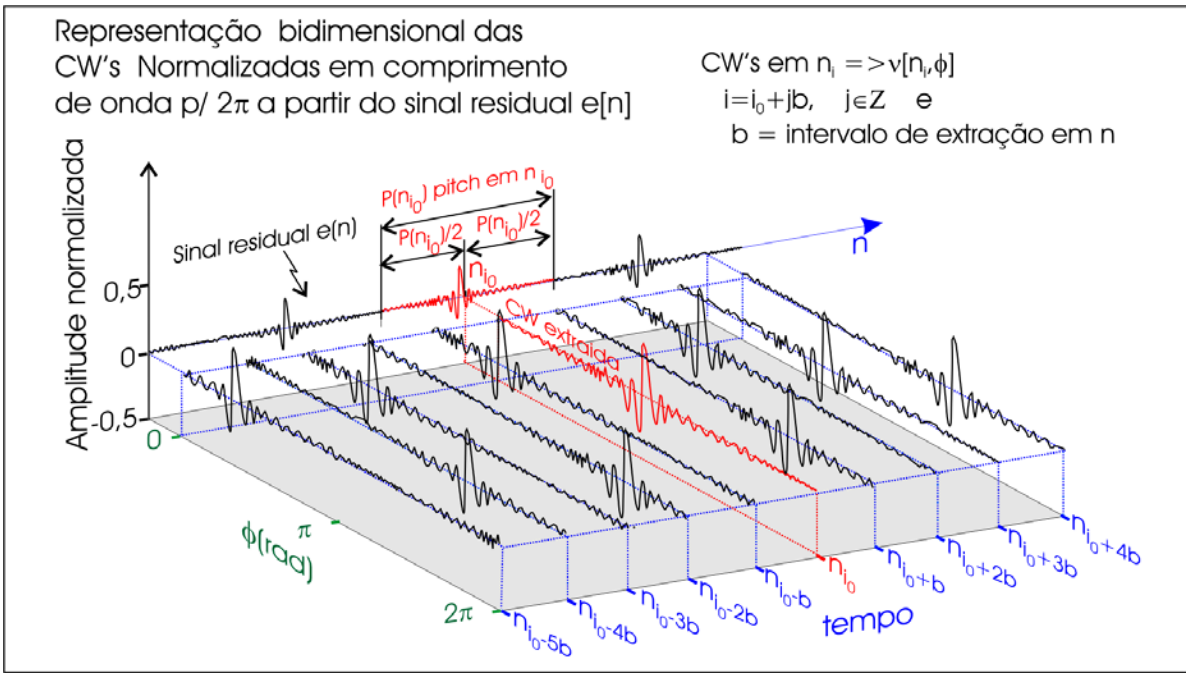

Figura 4.8 - Diagrama esquemático – Representação das formas de ondas características, CW's, em duas dimensões, normalizadas para comprimento em 2π => $v[n_i,\phi]$.

Considerações sobre a representação na DTFS (Série Discreta de Fourier no Tempo)

- As CW's, representadas no domínio do tempo por $v[n_i,m]$, Equação(4.4) ou $v[n_i,\phi]$, Equação(4.6) podem então ser representadas no domínio da freqüência pelos coeficientes de Fourier $\{A_k(n_i),B_k(n_i)\}$ referentes a cada CW, e calculados utilizando-se a Equação (4.3).

4.3.3 O estágio da análise

No esquema do codificador WI, Figura 4.4, a função do bloco de análise (processo C100), é converter o sinal da fala de entrada em formas de ondas características, CW's, bem como extrair os outros parâmetros ortogonais incluindo o pitch, os coeficientes LSFs e a potência.

A análise consiste em vários processos distintos que são esquematizados na Figura 4.9 onde o bloco de análise (processo C100) está expandido. O diagrama esquemático representa uma implementação típica WI. A descrição de cada um dos processos envolvidos é introduzida nas próximas subseções.

O sinal original da fala, s[n], na entrada (codificador) e o sinal da fala reconstruído, $\tilde{s}[n]$, na saída (decodificador) estão em um formato digital amostrado a uma freqüência de amostragem, Fs = 11,025 kHz. O sinal s[n] é dividido em quadros consecutivos $\ldots,(s[n]_1),(s[n]_2)\ldots,(s[n]_{l-1}),(s[n]_l),(s[n]_{l+1})\ldots$ com um comprimento L_q, onde $(s[n]_l)$ representa o quadro atual, $(s[n]_{l-1})$ o quadro anterior (ou passado) e $(s[n]_{l+1})$ o quadro

posterior ou em avanço. Também os quadros são divididos em sub-quadros consecutivos $(s[n])_{sq0}, (s[n])_{sq1}, \ldots, (s[n])_{sq(i-1)}, (s[n])_{sqi}, (s[n])_{sq(i+1)}, \ldots, (s[n])_{sq(Nsq)}$ com um comprimento L_{sq}, onde Nsq é o número de sub-quadros em um quadro do sinal de voz, $(s[n])_{sqi}$ representa o sub-quadro atual, $(s[n])_{sq(i-1)}$ o sub-quadro anterior e $(s[n])_{sq(i+1)}$ o sub-quadro posterior. Na implementação do codificador foram utilizados os parâmetros: L_q = 160 amostras (14,5 ms), L_{sq} = 20 amostras (1,8 ms) e Nsq = 8 sub-quadros. Alguns processos são realizados uma vez por quadro enquanto outros são realizados uma vez por sub-quadro.

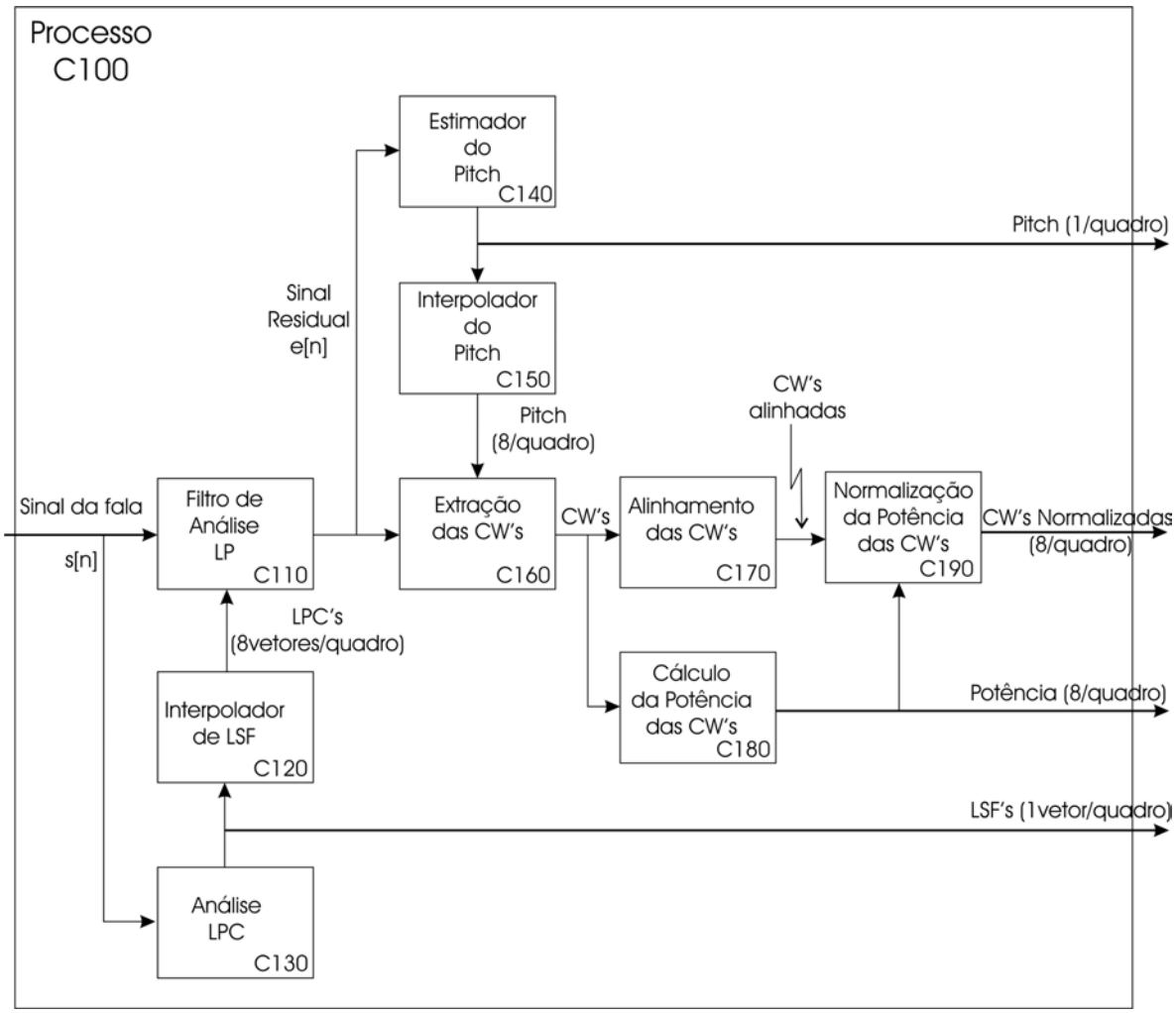

Figura 4.9 – Diagrama de blocos de análise expandido (processo C100) do codificador WI.

4.3.3.1 Análise com predição linear (análise LP)

A análise LP compreende os blocos C110, C120 e C130. O processamento começa pelo bloco C130 (Análise LPC) composto pelas operações esquematizadas na Figura 4.10: **(a)** pré-ênfase; **(b)** Enjanelamento do sinal (janela de Hamming); **(c)** cálculo dos coeficientes LPC (método da autocorrelação); **(d)** expansão da largura de faixa; e **(e)** conversão dos coeficientes

LPC para os coeficientes LSF (conforme o algoritmo descrito na seção 2.5.1.3.1 do capítulo 2 deste trabalho). A análise LP é realizada para estimar os coeficientes LPC correspondentes a um quadro, o quadro atual. No caso específico mostrado na Figura 4.10 o quadro atual corresponde ao quadro *l*, $(s[n]_l)$. Inicialmente o conjunto de amostras, $(s[n])_{Lwl}$, relacionadas com a análise do quadro atual é selecionado por uma janela de comprimento L_{Wl} = 240 amostras e enviado ao bloco C130. A janela é centrada no limite superior do quadro atual, envolvendo 120 amostras do quadro atual, $(s[n]_l)$, e 120 amostras do quadro posterior, $(s[n]_{l+1})$. Estas últimas 120 amostras correspondem a um atraso de 10,9 ms devido ao algoritmo. As amostras na janela, $(s[n])_{Lwl}$, são então, **(a)** pré-enfatizadas usando a Equação (2.76) resultando em $(s_{enf}[n])_{Lwl}$ (neste capítulo o fator de pré-ênfase λ na Equação (2.76) será trocado por α com o valor de 0,9). **(b)** Às amostras pré-enfatizadas da fala, $(s_{enf}[n])_{Lwl}$, aplica-se a janela de Hamming obtida pela Equação (2.21), resultando nas amostras $(s_w[n])_{Lwl}$ que são então utilizadas para, **(c)** calcular os coeficientes LPC $\{a_k\}_l$ utilizando-se o método da autocorrelação, com uma análise LP de 10ª ordem (p=10). Em **(d)** os coeficentes $\{a_k\}_l$ são submetidos a uma expansão da largura de faixa usando $\gamma = 0{,}98829$ conforme a Equação $\{a_k\}_{Exp(l)} = \{\gamma^k a_k\}_l \quad (k = 1, \cdots, p)$ relacionada na seção 2.5.2, o que corresponde a uma expansão de 30 Hz e resulta nos coeficientes $\{a_k\}_{Exp(l)}$ [14]. E finalmente, em **(e)** aos coeficientes $\{a_k\}_{Exp(l)}$ aplica-se o algoritmo para executar a conversão dos coeficientes LPC para os coeficientes LSF (seção 2.5.1.3.1), resultando então nos coeficientes $\{a_k\}_{LSF(l)}$ que representam o quadro atual, no caso o quadro *l*. O vetor $\{a_k\}_{LSF(l)}$ é então enviado para o próximo bloco, *Interpolador de LSF* (processo C120). As operações são repetidas quadro a quadro resultando nos vetores LSF's, …,, $\{a_k\}_{LSF(l)}$, $\{a_k\}_{LSF(l+1)}$, $\{a_k\}_{LSF(l+2)}$, …, que representam os coeficientes de predição linear de cada quadro analisado, …, $(s[n]_l), (s[n]_{l+1}), (s[n]_{l+2}), \ldots,$ respectivamente. Os vetores LSF's são enviados por quadro ao processo C120, o que representa uma taxa de atualização de 69 Hz, um vetor LSF por quadro a cada 14,5 ms.

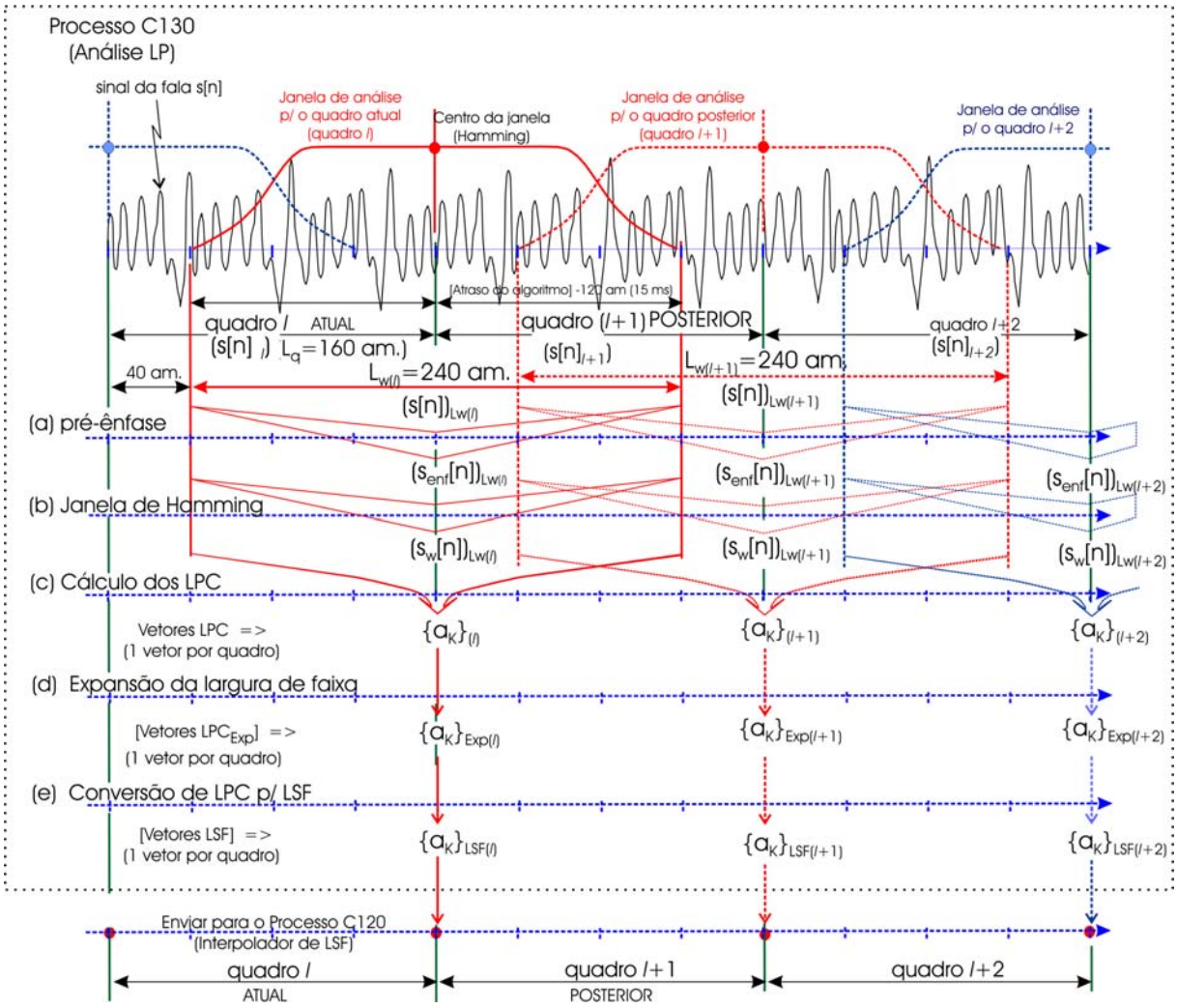

Figura 4.10 – Diagrama esquemático das operações no bloco análise LP (processo C130) do codificador WI.

No bloco *Interpolador de LSF* (processo C120), o processamento é composto pelas operações esquematizadas na Figura 4.11: **(a)** Interpolação de LSFs (sub-quadros); e **(b)** Conversão de coeficientes LSF para coeficientes LPC (sub-quadros), conforme a seção 2.5.1.3.2 do capítulo 2 deste trabalho. Na primeira operação, em **(a),** os vetores LSF entre dois quadros sucessivos, $\{a_k\}_{LSF(l-1)}$ $(quadro\ anterior)$ e $\{a_k\}_{LSF(l)}$ $(quadro\ atual)$, são interpolados à nível de sub-quadro, $\{a_k\}_{LSF(sq0)}$, $\{a_k\}_{LSF(sq1)},\ldots,\{a_k\}_{LSF(sqi)},\ldots,\{a_k\}_{LSF(sq7)}$, para garantir uma transição suave entre os parâmetros. Em seguida, em **(b)**, os coeficientes LSF (sub-quadros) são convertidos para LPC (sub-quadro), $\{a_k\}_{sq0}$, $\{a_k\}_{sq1},\ldots,\{a_k\}_{sqi},\ldots,\{a_k\}_{sq7}$, e enviados ao bloco *Filtro de análise LP* (processo C110).

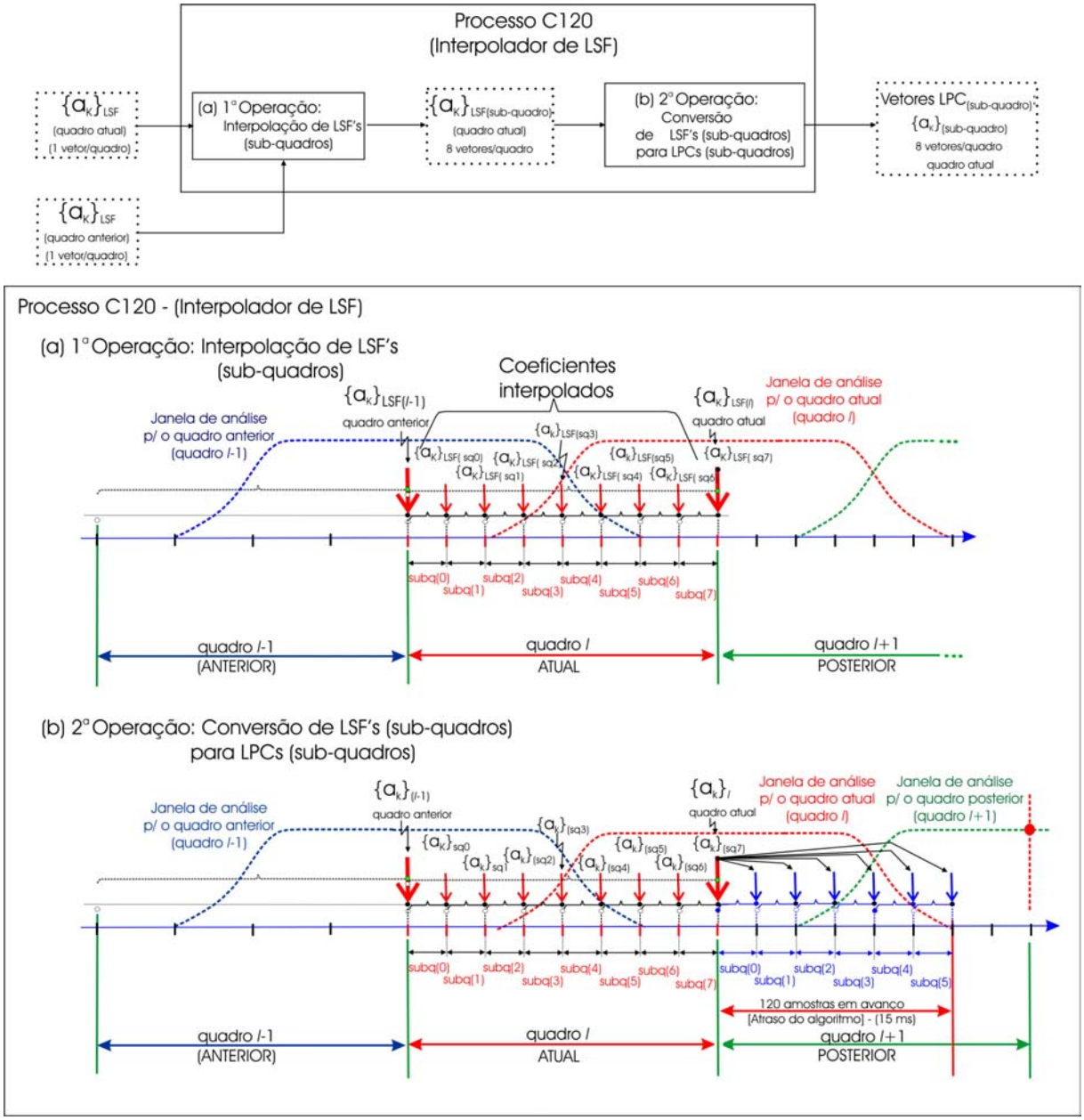

Figura 4.11 – Diagrama esquemático das operações no bloco Interpolador de LSF's (processo C120) do codificador WI.

No bloco *Filtro de análise LP* (processo C110), o processamento é composto pelas operações esquematizadas na Figura 4.12-a e b: **(a)** Cálculo do sinal residual ($e[n]_l$) para o quadro atual ($s[n]_l$); e **(b)** Cálculo do sinal residual ($e[n]_{av}$) para as amostras em avanço no quadro posterior ($s[n]_{l+1}$). Estas amostras ($e[n]_{av}$), do sinal residual em avanço também são utilizadas na estimação do pitch no processo C140. Inicialmente, em **(a),** o *Filtro de análise LP* recebe $(s[n])_{mem}$, as últimas p amostras do quadro anterior ($s[n]_{l-1}$),para compor a memória inicial do filtro, e ($s[n]_l$), as (Lq = 160) amostras relativas ao quadro atual que serão processadas em conjuntos consecutivos com 20 amostras, ou seja, em sub-quadros com (Lsq = 20 amostras). A seqüência de amostras nos sub-quadros, denominadas por

$(s[n])_{sq0}, (s[n])_{sq1}, \ldots, (s[n])_{sqi,\ldots,} (s[n])_{sq7}$, são pré enfatizadas usando um fator $\alpha = 0{,}1$ $(\lambda = \alpha)$ na Equação (2.76) resultando em $(s_{enf}[n])_{sq0}, (s_{enf}[n])_{sq1}, \ldots, (s_{enf}[n])_{sqi,\ldots,} (s_{enf}[n])_{sq7}$. O *Filtro de análise LP* recebe também os coeficientes LPC (por sub-quadro), $\{a_k\}_{sq0}$, $\{a_k\}_{sq1},\ldots,\{a_k\}_{sqi},\ldots,\{a_k\}_{sq7}$ (onde $1 \le k \le p$) que são convertidos para os coeficientes do filtro do erro de predição $\{ec_k\}_{sq0}, \{ec_k\}_{sq1},\ldots,\{ec_k\}_{sqi},\ldots,\{ec_k\}_{sq7}$, onde $0 \le k \le p$, de forma que $\{ec_0 = 1, ec_1 = -a_1, ec_2 = -a_2, \cdots, ec_p = -a_p\}_{sqi}$. Para cada sub-quadro $(s_{enf}[n])_{sqi}$ com o seu respectivo vetor de coeficientes do filtro do erro de predição $\{ec_k\}_{sqi}$ (por sub-quadro), executa-se a filtragem, utilizando-se a Equação (2.46), obtendo-se o respectivo sub-quadro do sinal residual $(e[n])_{sqi}$. Após o cálculo de todos os sub-quadros $(e[n])_{sq0}, (e[n])_{sq1}, \ldots, (e[n])_{sqi}, \ldots, (e[n])_{sq7}$ que compõem o quadro atual $(e[n]_l)$, executa-se em **(b)** operações análogas às realizadas em **(a),** exceto que o vetor de coeficientes do filtro de predição de erro $\{ec_k\}_{sq7}$ é utilizado para a filtragem de todos os sub-quadros das $(L_a = 120)$ amostras em avanço, $(s[n]_{av})$, definidos como $(s[n]_{av})_{sq0}, (s[n]_{av})_{sq1}, \ldots, (s[n]_{av})_{sqi}, \ldots, (s[n]_{av})_{sq5}$ na obtenção dos sub-quadros $(e[n]_{av})_{sq0}, (e[n]_{av})_{sq1}, \ldots, (e[n]_{av})_{sqi}, \ldots, (e[n]_{av})_{sq5}$ que compõem o sinal residual em avanço, $(e[n]_{av})$. Assim na saída do filtro as seqüências $(e[n]_l)$ e $(e[n]_{av})$ que compõem o sinal residual $(e[n])$ relativo ao quadro atual ficam disponíveis para a estimação do pitch do quadro atual no processo C140 (*Estimador do Pitch*). Os sinais residuais, $(e[n])$ relativo ao quadro atual e $(e[n]_{l-1})$ do quadro anterior calculado em uma etapa anterior, também ficam disponíveis para a extração das CW's no processo C160 (*Extração das CW's*), ou seja, as últimas *delta1* (amostras) em $(e[n]_{l-1})$ + $(e[n]_l)$ + as primeiras *delta2* (amostras) em $(e[n]_{av})$, como é mostrado a Figura. 4.17.

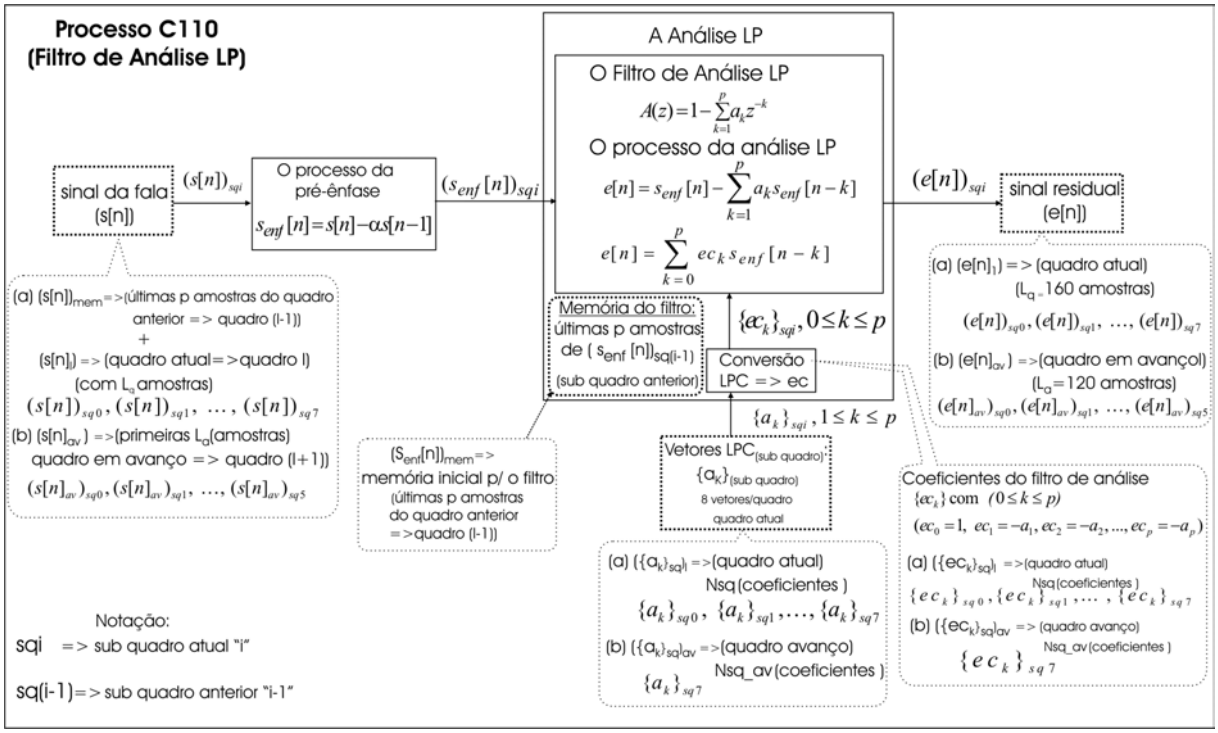

Figura 4.12-a – Diagrama esquemático das operações no bloco do Filtro de Análise LP (processo C110) do codificador WI – equação e notação dos parâmetros.

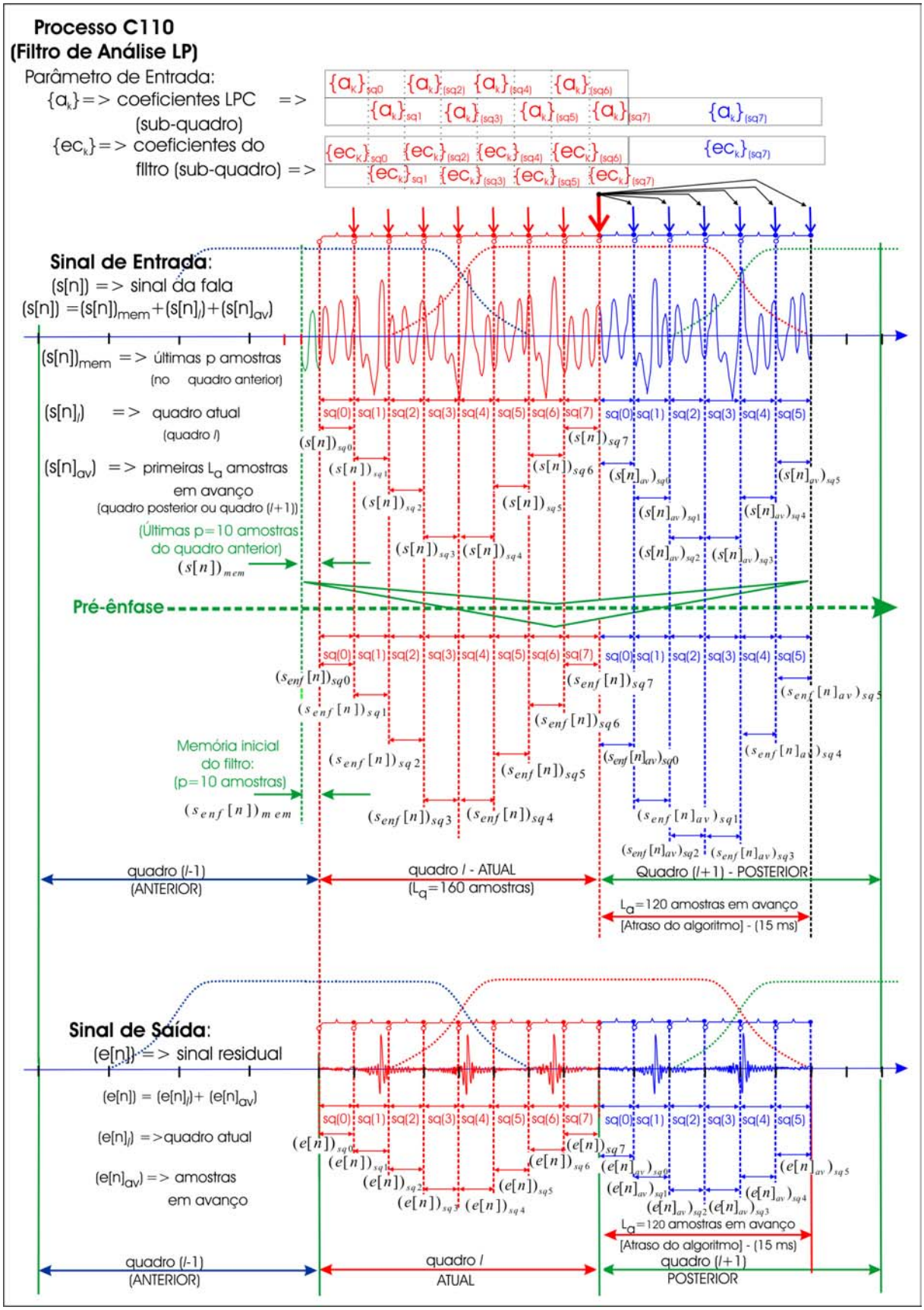

Figura 4.12-b – Diagrama esquemático das operações no bloco do Filtro de Análise LP (processo C110) do codificador WI – Visualização esquemática dos sinais de entrada e de saída do filtro.

4.3.3.2 Estimação do pitch

O pitch é uma propriedade fundamental da fala sonora, pois descreve a freqüência fundamental da fala que é a freqüência de vibração das cordas vocais que produz os sons sonoros da fala [66]. A evolução do pitch é estimada no codificador WI, pois o modelo de codificação WI se baseia nas informações do período do pitch. No codificador o pitch é estimado por quadro, $\text{Pitch}_{(\text{por quadro})}$, e interpolado por sub-quadro, $\text{Pitch}_{(\text{por sub-quadro})}$. No decodificador ocorre a interpolação do pitch por sub-quadro, $\text{Pitch}_{(\text{por sub-quadro})}$, e por amostra, $\text{Pitch}_{(\text{por amostra})}$. A precisão do estimador do período de pitch é de fundamental importância para o desempenho do codificador [14]. O pitch pode ser estimado, por exemplo, a partir do sinal da fala ou do sinal residual. Vários algoritmos diferentes têm sido desenvolvidos para este propósito nas últimas décadas, cada um tendo seus pontos fortes e pontos fracos. O algoritmo pode ser baseado na determinação dos picos no ciclo do pitch a partir do sinal residual ou na determinação do deslocamento que maximiza a autocorrelação para um quadro de amostras da fala [66].

Neste trabalho a estimação do período do pitch é realizada através do *Estimador de Pitch* - processo C140 que pertence a última categoria, onde é utilizada a função autocorrelação normalizada (FACN) ou a função correlação cruzada normalizada (FCCN). O processo C140 (*Estimador do Pitch*) recebe o sinal residual $(e[n])$ relativo ao quadro atual e de acordo com uma das quatro janelas (de pesquisa do pitch) que são descritas a seguir, aplica-se a função autocorrelação normalizada (FACN),

$$FACN(d) = \frac{\sum_{m=0}^{N-d-1} e[m] * e[m+d]}{\sqrt{\sum_{m=0}^{N-d-1} e^2[m]} \sqrt{\sum_{m=0}^{N-d-1} e^2[m+d]}} \tag{4.7}$$

ou a função correlação cruzada normalizada (FCCN),

$$FCCN(d) = \frac{\sum_{m=0}^{N-1} e[m] * e[m+d]}{\sqrt{\sum_{m=0}^{N-1} e^2[m]} \sqrt{\sum_{m=0}^{N-1} e^2[m+d]}} \tag{4.8}$$

onde N é número de amostras da janela envolvida no cálculo (janela fixa), d é o deslocamento da janela deslizante (em número de amostras) em relação à janela fixa, $FACN(d)$ é a autocorrelação normalizada para a janela deslocada de d amostras e $FCCN(d)$ é a correlação cruzada normalizada para a janela deslocada de d amostras. O objetivo de cada uma das janelas de pesquisa do pitch é determinar o deslocamento d (*o valor pitch em amostras*) que maximiza a autocorrelação ou a correlação cruzada para um quadro

de amostras da fala. A pesquisa do pitch (*número inteiro*) é realizada na faixa $P_{mín} \le d \le P_{máx}$, onde $P_{mín} = 20\ amostras$ e $P_{máx} = 120\ amostras$, são limites típicos para a fala amostrada em 11,025 kHz, e foram definidos em função dos limites da produção da fala humana e que também permite uma codificação para o pitch com sete bits sem erro de quantização.

A estimação do pitch, no processo C140, é composta pelas operações esquematizadas na Figura 4.13-a e b: **(a)** processo C141 - **Janela$_1$**; **(b)** processo C142 - **Janela$_{1R}$**; **(c)** processo C143 - **Janela$_2$**; **(d)** processo C144 - **Janela$_3$**; e **(e)** processo C145 – **Regras de decisão**. Na primeira operação, em **(a),** processo C141, **a Janela$_1$** enfatiza as mostras do quadro atual envolvendo todas as amostras do quadro ($N = L_q = 160$), as quais são processadas utilizando-se a FCCN, Equação (4.8), tendo como resultado o deslocamento d_1 que corresponde à FCCN máxima para o intervalo $P_{mín} \le d \le P_{máx}$, denotada por $R_{máx1}$. Em seguida, em **(b)**, o processo C142 - **Janela$_{1R}$** recebe o resultado do processo C141 - **Janela$_1$,** d_1 e $R_{máx1}$ que são usados para a determinação mais refinada do pitch, considerando a **Janela$_{1R}$** com comprimento $N = 2 * d_1$ e pesquisando o pitch na faixa $d_1 - 10 \le d \le d_1 + 10$. As amostras são processadas utilizando-se a FCCN, Equação (4.8), tendo como resultado o deslocamento d_{1R} que corresponde à FCCN máxima, $R_{máx1R}$. Em **(c)**, processo C143 a **Janela$_2$** enfatiza as mostras em avanço do quadro posterior, envolvendo as últimas 40 amostras do quadro atual mais as L_a =120 amostras em avanço, em um total de $N = 160$. As amostras são processadas utilizando-se a FCCN, com a janela se deslocando no sentido inverso com o intervalo $P_{mín} \le d \le P_{máx}$, tendo como resultado d_2 que corresponde à FCCN máxima, $R_{máx2}$. Em **(d)**, processo C144, a **Janela$_3$** enfatiza as amostras na fronteira compreendendo as últimas 80 amostras do quadro atual e as primeiras 80 amostras em avanço no quadro posterior, $N = 160$. As amostras são processadas utilizando-se a FACN, Equação (4.7), com a janela se deslocando no intervalo $P_{mín} \le d \le P_{máx}$, tendo como resultado d_3 que corresponde à FACN máxima, $R_{máx3}$. Finalizando, em **(e),** processo C145 – **Regras de decisão,** chega-se ao pitch estimado para o quadro atual. Neste processo foi adotado um conjunto de regras para escolher o melhor pitch estimado considerando a importância da posição das janelas e os cálculos efetuados para cada uma delas como: **Janela$_1$** (d_1, $R_{máx1}$), **Janela$_{1R}$** (d_{1R}, $R_{máx1R}$), **Janela$_2$** (d_2, $R_{máx2}$) e a **Janela$_3$** (d_3, $R_{máx3}$). Foram considerados também alguns limites verificados no levantamento estatístico do sinal na comparação entre os valores do pitch obtidos utilizando-se os cálculos nas janelas e o valor do pitch obtido no cálculo manual para algumas expressões da fala. Assim a lógica da escolha é descrita no quadro a seguir:

(Início){
Se ($R_{máx1R} > R_{máx2} + 0{,}21$){ // enfatiza as //Amostras em avanço
$d_{ótimo} = d_{1R}$;
$R_{ótimo} = R_{máx1R}$;
}
Se não, então{
$d_{ótimo} = d_2$;
$R_{ótimo} = R_{máx2}$;
}
	Se ($R_{máx3} > R_{ótimo} + 0{,}21$){ //enfatiza o método // FCCN
	Se (abs(d_3-$d_{ótimo}$) < 2)
	Se ($d_3 < d_{ótimo}$) //segue o menor
	$d_{ótimo} = d_3$;
	Se (abs(d_3-$d_{ótimo}$) >1)
	Se ($d_3 > d_{ótimo}$) //segue o maior
	$d_{ótimo} = d_3$;
	}
	Se ($R_{ótimo} \leq 0{,}31$) // lim. entre sonoro/surdo
	$d_{otimo} = P_{mín}$;
	Faz $Pitch_{(quadro\ atual)} = d_{ótimo}$;
	}(Fim)

O valor do $pitch_{(quadro\ atual)}$ ($pitch_l$ ou $P(n_l)$ em número de amostras) mais o valor do $pitch_{(quadro\ anterior)}$, $pitch_{(l-1)}$ ou $P(n_{l-1})$ ficam disponível para o processo C150 (Interpolador do pitch) e também para serem quantizados e/ou enviados ao decodificador.

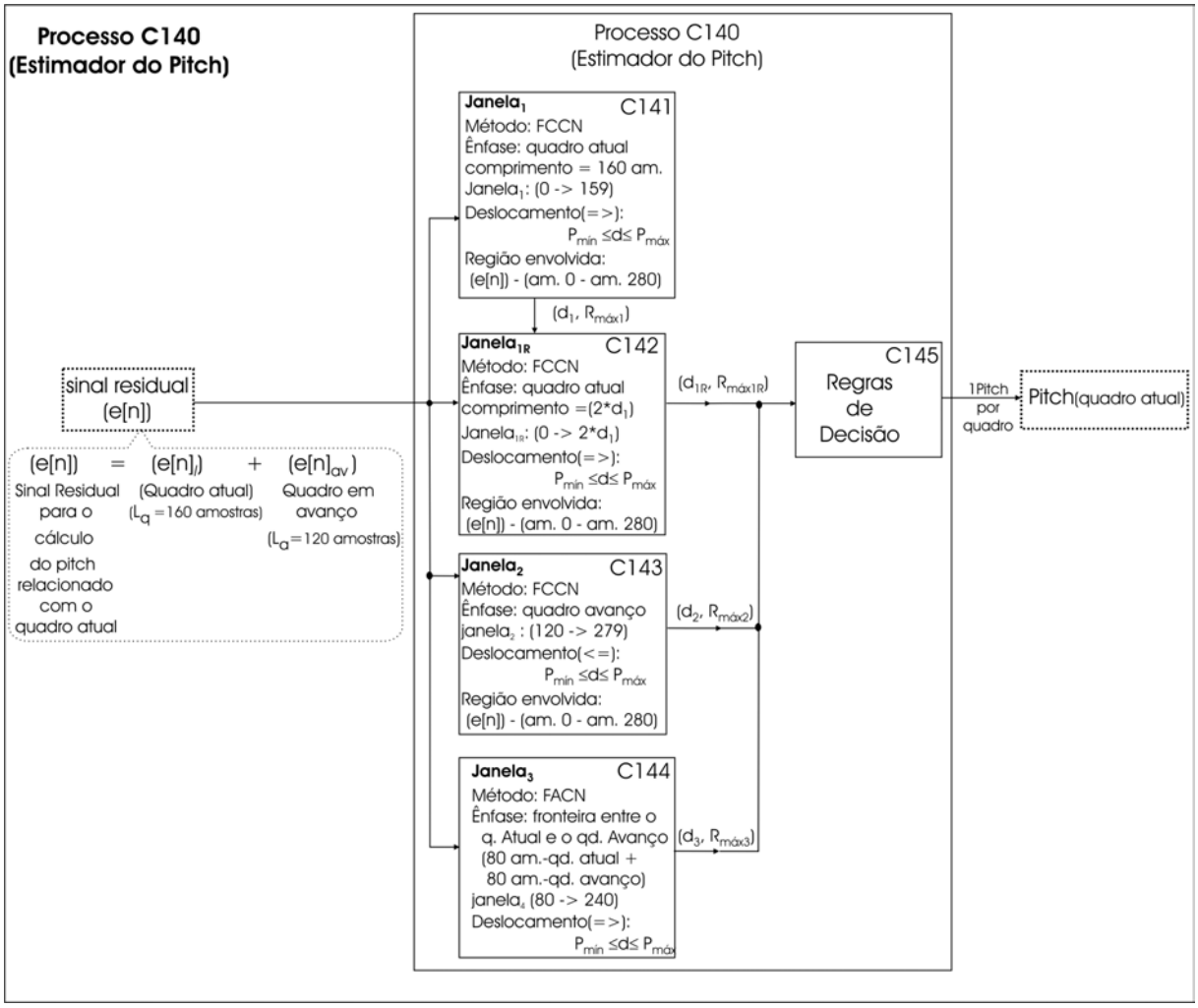

Figura 4.13-a – Diagrama esquemático das operações no bloco Estimador do Pitch (processo C140) do codificador WI – Resumo das características em cada processo.

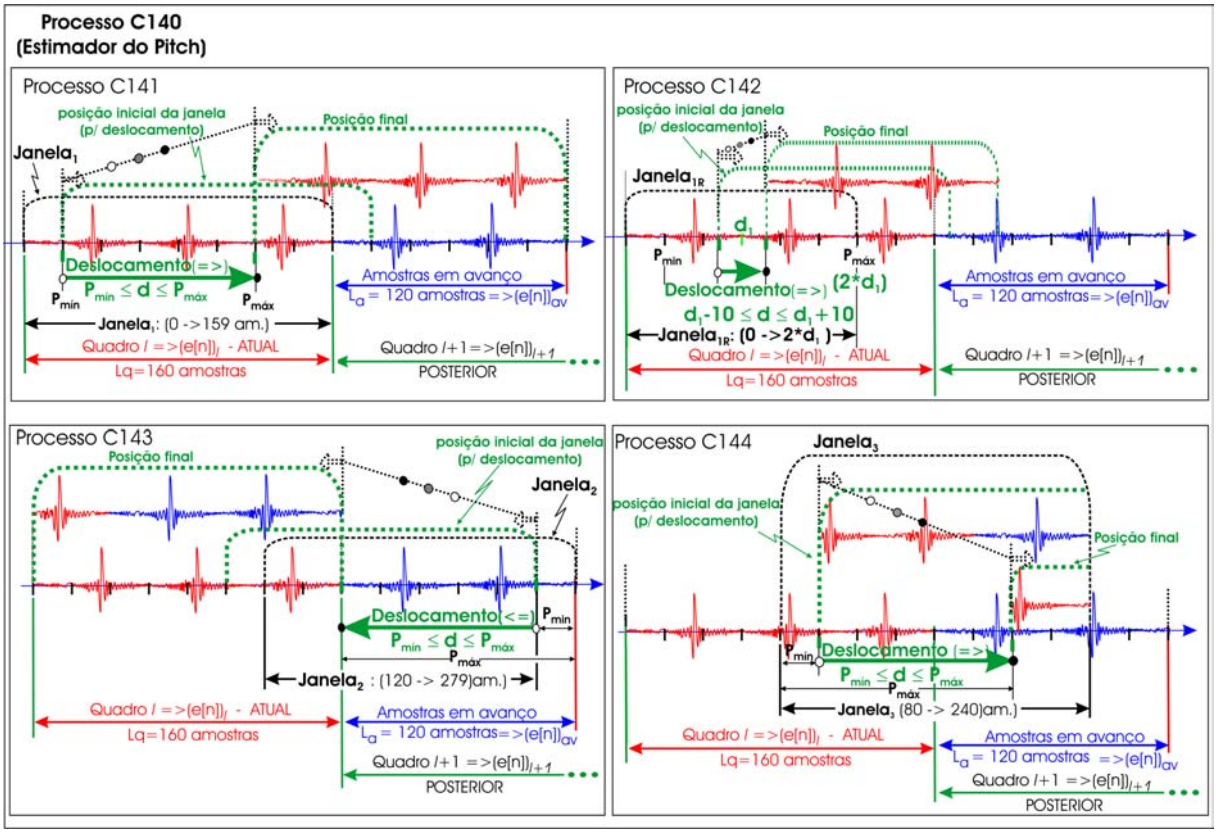

Figura 4.13-b – Diagrama esquemático das operações no bloco Estimador do Pitch (processo C140) do codificador WI – Visualização esquemática do sinal e do processamento do sinal residual na determinação do deslocamento d ($pitch_{quadro\ atual}$ - $pitch_l$ => $P(n_l)$ em número de amostras).

4.3.3.3 Interpolação do Pitch

Na técnica WI o comprimento de cada CW é função do período do pitch que está relacionado com n_i, ponto de extração da CW. No processo C160 do codificador as CW's são extraídas a uma taxa de 1CW por sub-quadro. Portanto é necessária a interpolação do valor do pitch por sub-quadro, pois o pitch foi estimado por quadro no processo C140. A interpolação linear básica tem sido considerada suficiente para a técnica WI, mas foi observado que existem situações onde este método não pode ser usado, pois poderá resultar em distorções audíveis na fala reconstruída [14]. Tal situação ocorre, por exemplo, quando o pitch é duplicado ou dividido ao meio [67]. Um método que tem sido usado para a interpolação do pitch em implementações WI é uma combinação da interpolação linear com a interpolação com degrau.

No processo C150 – *Interpolador do Pitch*, a interpolação é executada utilizando-se a interpolação linear básica ou a interpolação linear com degrau, como mostrado na Figura 4.14. O *Interpolador do Pitch recebe os* valores do $pitch_{(quadro\ atual)}$, $P(n_l)$ em número de amostras, e o valor do $pitch_{(quadro\ anterior)}$, $P(n_{l-1})$, executa a *interpolação linear básica*

quando $|P(n_l) - P(n_{l-1})| < (Limiar = 20)|$ e a interpolação linear com degrau quando $|P(n_l) - P(n_{l-1})| \geq (Limiar = 20)$. Assim, considerando que i seja substituído consecutivamente por $\{sq0, sq1, \ldots, sq7\}$, então o valor do pitch interpolado em n_i, $P(n_i)$, pode ser calculado usando-se as Equações 4.9 à 4.13:

Caso1 (*Interpolação linear básica*):

$$P(n_i) = \frac{[(n_l - n_i)P(n_{l-1}) + (n_i - n_{l-1})P(n_l)]}{(n_l - n_{l-1})} \qquad n_{l-1} < n_i \leq n_l \tag{4.9}$$

Caso2 (*Interpolação linear com degrau*):

(a) Para $P(n_l) > P(n_{l-1}) => C = P(n_l)/P(n_{l-1})$, onde C é arredondado para o inteiro mais próximo,

Patamar1 (*Interpolação linear)*:

$$P(n_i) = \frac{C(n_l - n_i)P(n_{l-1}) + (n_i - n_{l-1})P(n_l)}{C(n_l - n_{l-1})} \qquad n_{l-1} \leq n_i < (n_l + n_{l-1})/2 \tag{4.10}$$

Patamar2 (*Interpolação linear)*:

$$P(n_i) = \frac{C(n_l - n_i)P(n_{l-1}) + (n_i - n_{l-1})P(n_l)}{(n_l - n_{l-1})} \qquad [(n_l + n_{l-1})/2] \leq n_i < n_l \tag{4.11}$$

(b) Para $P(n_{l-1}) > P(n_l) => C_1 = P(n_{l-1})/P(n_l)$, onde C_1 é arredondado para o inteiro mais próximo,

Patamar1 (*Interpolação linear)*:

$$P(n_i) = \frac{(n_l - n_i)P(n_{l-1}) + C_1(n_i - n_{l-1})P(n_l)}{(n_l - n_{l-1})} \qquad n_{l-1} \leq n_i < (n_l + n_{l-1})/2 \tag{4.12}$$

Patamar2 (*Interpolação linear)*:

$$P(n_i) = \frac{(n_l - n_i)P(n_{l-1}) + C_1(n_i - n_{l-1})P(n_l)]}{C_1(n_l - n_{l-1})} \qquad [(n_l + n_{l-1})/2] \leq n_i < n_l \tag{4.13}$$

Após as operações o processo C150 – *Interpolador do Pitch* deixa disponíveis os valores de pitch $\{P(n_{sq0}), P(n_{sq1}), \ldots, P(n_{sq7})\}$ para extração das CW's do quadro atual utilizando-se o processo C160.

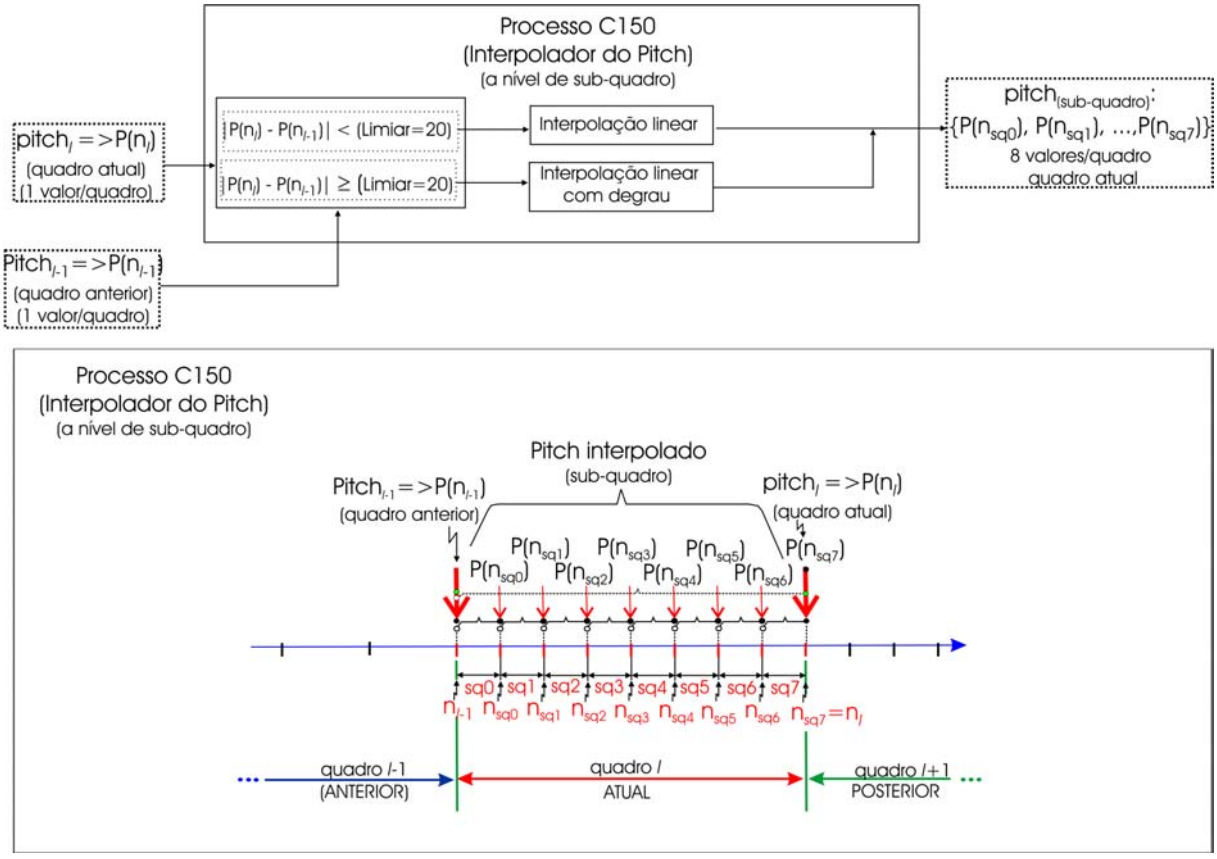

Figura 4.14 – Diagrama esquemático das operações no bloco Interpolador do Pitch (processo C150) do codificador WI.

4.3.3.4 Extração das CW's

Após os períodos de pitch por sub-quadros, $\{P(n_{sq0}), P(n_{sq1}), \ldots, P(n_{sq7})\}$, para o quadro atual terem sido interpolados, as CW's, formas de ondas características, podem então ser extraídas no processo C160 – *Extração das CW's,* esquematizada na Figura 4.15. Além das amostras do sinal residual no quadro atual, $(e[n]_l)$, também ficam disponíveis para auxiliar no processo de extração, as amostras do sinal residual que compreendem as últimas *delta1* amostras do quadro anterior $(e[n]_{l-1})$, ou as amostras passadas, e as primeiras *delta2* amostras no quadro posterior $(e[n]_{av})$, como mostrado na Figura 4.17. Um segmento com $P(n_{sqi})$ amostras (uma CW) é extraída do sinal residual na região em torno de cada instante n_{sqi} $(i = 0,1,\ldots,Nsq-1)$, que representa a posição da última amostra de cada sub-quadro, no quadro atual.

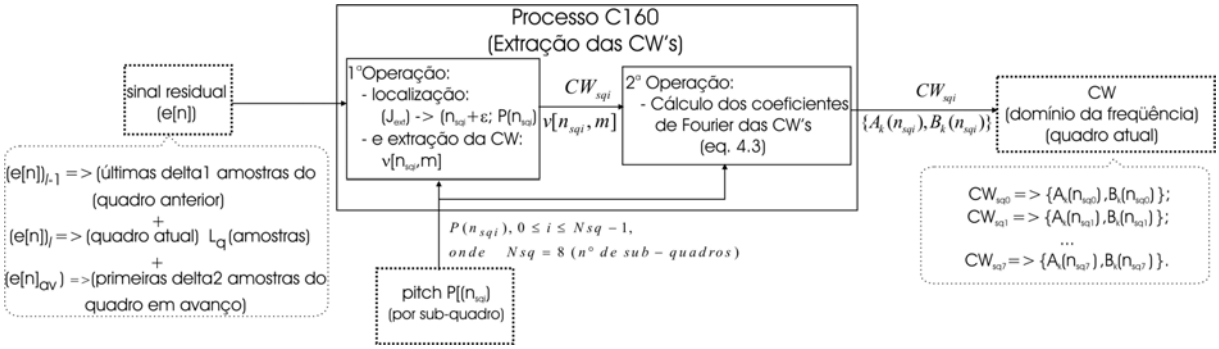

Figura 4.15 – Diagrama esquemático do bloco *extração das CW's* (processo C160) no codificador WI.

Cada CW é representada como foi definido na Equação (4.1) e ilustrada na Figura 4.6. A definição da posição do segmento no sinal residual para a extração da CW é realizada (no domínio do tempo) utilizando-se a Equação,

$$v[n_{sqi}, m] = e[n_{sqi} + \varepsilon - \frac{P(n_{sqi})}{2} + m] \quad \text{onde} \quad 0 \le m < P(n_{sqi}), \tag{4.14}$$

(CW extraída do sinal residual)

sendo representada (no domínio da freqüência) em DTFS utilizando-se a Equação (4.4), ou ainda (no domínio do tempo) pela Equação,

$$v[n_{sqi}, \phi] = e[n_{sqi} + \varepsilon - \frac{P(n_{sqi})}{2} + \frac{\phi P(n_{sqi})}{2\pi}] \tag{4.15}$$

(CW normalizada - comprimento igual a 2π)

também representada (no domínio da freqüência) em DTFS utilizando-se a Equação (4.6). O parâmetro ε é um pequeno ajuste na posição do segmento que representam a CW, que permite minimizar as descontinuidades causadas pela extensão periódica

$$v[n_{sqi}, \phi] = v[n_{sqi}, \phi + 2\pi\eta], \qquad \eta \in Z \tag{4.16}$$

que é suposta durante a conversão para o domínio DTFS ao utilizar a representação utilizando-se a Equação (4.4) ou a Equação (4.6). Assim na extração, procura-se a melhor posição do segmento para que a energia seja mínima nas regiões em torno dos extremos das CW's. Na realização desta operação é utilizada uma janela, denominada de *janela de extração*, que é esquematizada na Figura 4.16. No início da extração uma janela denominada *janela inicial* ($J_{inicial}$) com tamanho $P(n_{sqi})$ é posicionada com seu centro coincidindo com a posição n_{sqi}. Em cada extremo da *janela inicial* é definida uma *janela de verificação da energia* com comprimento δ e centro coincidindo com o extremo da *janela inicial*, a *janela*

de verificação da energia (extremo esquerdo), J_{ee}, e a *janela de verificação da energia (extremo direito),* J_{ed}. O conjunto, *janela inicial* mais as duas *janelas de energia* que permanecem ligadas aos extremos da *janela inicial*, é permitido se deslocar na faixa $-\varepsilon_{máx} \le \varepsilon' \le \varepsilon_{máx}$ em torno do ponto de extração da janela inicial. Para cada posição ε', a energia limitada (E_{limitada}) é calculada como a soma da energia verificada pelas duas janelas, E_{jee} e E_{jed}. A posição ε que corresponde à posição das *janelas de energia* com energia limitada mínima define a posição final de extração, $n_{sqi} + \varepsilon$ que é o centro da *janela de extração*, J_{Ext}. A posição final de extração, $n_{sqi} + \varepsilon$, mais o pitch do sub-quadro i, $P(n_{sqi})$, definem a CW_{sqi} a ser extraída, relacionada à posição regular de extração n_{sqi}.

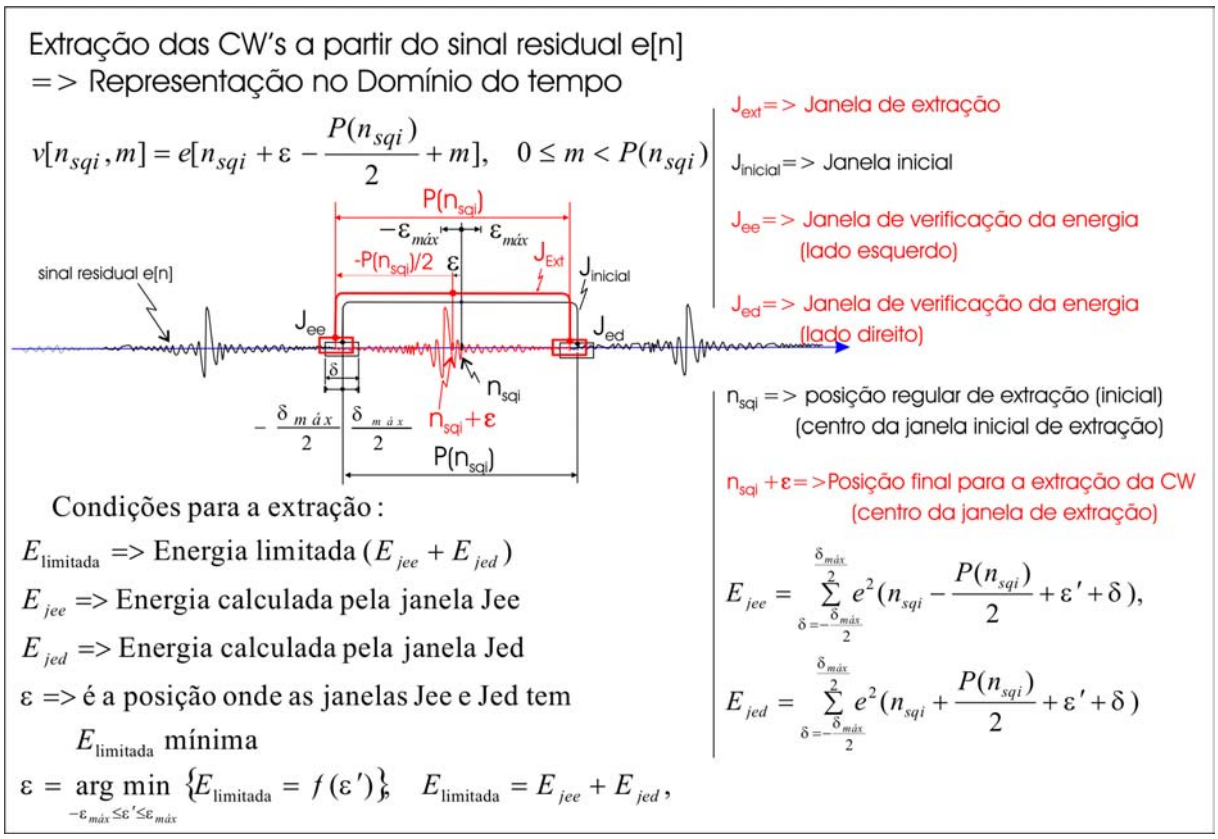

Figura 4.16 – Diagrama esquemático para a operação de extração – Janela de extração no (processo C160) do codificador WI.

Observações para a extração das CW's

- A operação de extração é um processo executado por quadro, para o quadro atual (uma extração por sub-quadro). Mas, devido ao método de extração descrito acima, são necessárias algumas amostras passadas, do *quadro anterior – delta1* amostras, e algumas amostras futuras, do *quadro posterior – delta2* amostras, como mostrado na Figura 4.17.

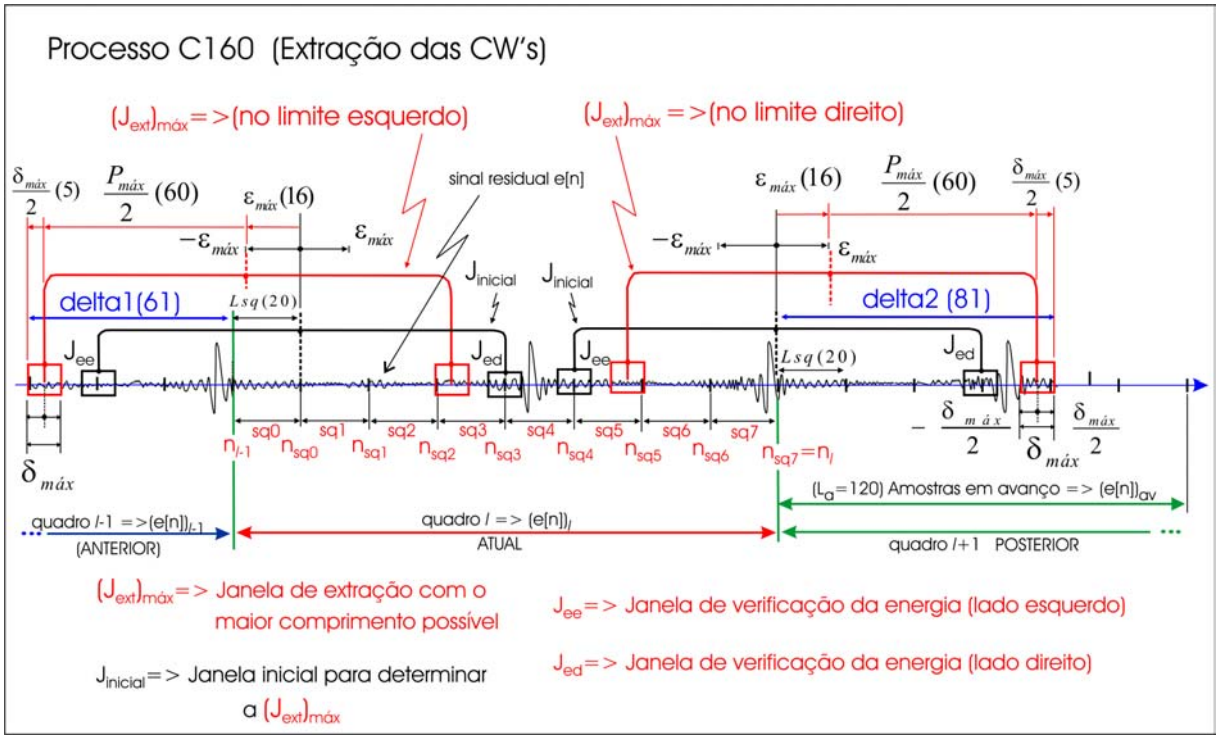

Figura 4.17 – Diagrama esquemático – Amostras necessárias para a operação de extração das CW's. Detalhe das posições da janela de extração nos extremos à esquerda e à direita na definição das amostras necessárias para o processo C160 – *Extração das CW's*.

As *delta1* amostras correspondem às últimas amostras no quadro anterior, assim definidas:

$$delta1 = \varepsilon_{máx} + \frac{P_{máx}(n_{sqi})}{2} + \frac{\delta_{máx}}{2} - L_{sq} \,. \tag{4.17}$$

As *delta2* amostras correspondem às primeiras amostras no quadro posterior, assim definidas:

$$delta2 = \varepsilon_{máx} + \frac{P_{máx}(n_{sqi})}{2} + \frac{\delta_{máx}}{2} \,. \tag{4.18}$$

Neste trabalho foram adotados como valores padrões, $\varepsilon_{máx} = 16$; $P_{máx}(n_{sqi}) = 120$; $\delta_{máx} = 10$ e $L_{sq} = 20$ que demonstraram bons resultados a partir de [14]. Assim $delta1 = 61$ amostras e $delta2 = 81$ amostras.

Após a extração, cada CW_{sqi} que fica representada no domínio do tempo por $v[n_{sqi}, m]$, Equação (4.14), é então transformada para o domínio da freqüência utilizando-se a Equação (4.3), onde são calculados os coeficientes de Fourier $\{A_k(n_{sqi}), B_k(n_{sqi})\}$, com

$0 \le i < Nsq$ ($N_{sq} = 8$, N_{sq} - número de sub-quadros). Assim as CW's utilizando-se os coeficientes de Fourier,

$$CW_{sq0} \Leftrightarrow \{A_k(n_{sq0}), B_k(n_{sq0})\};\quad CW_{sq1} \Leftrightarrow \{A_k(n_{sq1}), B_k(n_{sq1})\};\quad \cdots$$
$$\cdots;\ CW_{sq7} \Leftrightarrow \{A_k(n_{sq7}), B_k(n_{sq7})\}'$$

ficam disponíveis para os processos C170 (*Alinhamento das CW's*) e C180 (*Cálculo da Potência das CW's*), conforme é mostrado no esquema da Figura 4.9.

4.3.3.5 Alinhamento das CW's

O alinhamento entre duas CW's sucessivas, $CW_{anterior}$ e CW_{atual}, é realizado utilizando-se o deslocamento circular da forma de onda da CW_{atual}, ao longo do eixo de fase ϕ procurando a máxima correlação com a $CW_{anterior}$. Assim, um deslocamento de fase τ é selecionado tal que as características principais e similares nas duas ondas apareçam nos mesmos valores de fase. Esta operação é executada para minimizar as variações abruptas ou as descontinuidades na direção do eixo dos tempos n_i (eixo de evolução das formas de ondas), de forma que a superfície $u(n_i,\phi)$ fique suavizada com as CW's alinhadas.

Assim a CW atual extraída $v[n_{sqi}, \phi]$ é alinhada com a CW anterior $u[n_{sq(i-1)}, \phi]$ pela introdução do deslocamento de fase $\tau(n_{sqi})$ que maximiza a correlação cruzada entre as sucessivas formas de onda, onde a CW anterior $v[n_{sq(i-1)}, \phi]$ foi previamente alinhada e representada por $u[n_{sq(i-1)}, \phi]$. A CW atual alinhada pode, portanto, ser obtida usando-se:

$$u[n_{sqi}, \phi] = v[n_{sqi}, \phi + \tau(n_{sqi})] \tag{4.19}$$

onde:

$$\tau(n_{sqi}) = \arg \max_{\tau'} \int_0^{2\pi} v[n_{sqi}, \phi + \tau'] u[n_{sq(i-1)}, \phi] d\phi\,. \tag{4.20}$$

Assim $u[n_{sqi}, \phi]$ é a CW $v[n_{sqi}, \phi]$ após o alinhamento com a CW anterior $u[n_{sq(i-1)}, \phi]$ já alinhada. O alinhamento pode ser convenientemente executado no domínio da freqüência usando a representação em DTFS. Em [14] é mostrado que um deslocamento circular $T\,(0 \le T < P)$, em uma CW, no domínio do tempo, é equivalente a adicionar uma fase linear $\tau = \frac{2\pi T}{P}$ ($(0 \le \tau < 2\pi)$ ou deslocamento de fase) aos coeficientes DTFS da CW no domínio da freqüência. Assim o deslocamento de fase $\tau(n_{sqi})$, requerido para alinhar a CW atual com a CW anterior, considerando as duas CW's com a mesma dimensão, isto é,

$P(n_{sqi}) = P(n_{sq(i-1)})$ e $M = \frac{P(n_{sqi})}{2} = \frac{P(n_{sq(i-1)})}{2}$, é obtido pela maximização do seguinte critério, denominado de critério de alinhamento:

$$\tau(n_{sqi}) = \underset{0 \le \tau' < 2\pi}{\arg máx} \sum_{k=1}^{M} \{[A_k(n_{sq(i-1)})A_k(n_{sqi}) + B_k(n_{sq(i-1)})B_k(n_{sqi})]\cos(k\tau') + [B_k(n_{sq(i-1)})A_k(n_{sqi}) - B_k(n_{sqi})A_k(n_{sq(i-1)})]\text{sen}(k\tau')\} \quad (4.21)$$

O lado direito da Equação (4.21) é a correlação cruzada entre as CW's expressa em termo dos coeficientes da DTFS. O deslocamento de fase $\tau(n_{sqi})$ calculado é então aplicado para alinhar a CW atual, onde seus novos coeficientes DTFS são obtidos por:

$$\left.\begin{aligned} A'_k(n_{sqi}) &= A_k(n_{sqi})\cos(k\tau(n_{sqi})) - B_k(n_{sqi})\text{sen}(k\tau(n_{sqi})) \\ B'_k(n_{sqi}) &= A_k(n_{sqi})\text{sen}(k\tau(n_{sqi})) - B_k(n_{sqi})\cos(k\tau(n_{sqi})) \end{aligned}\right\} \text{ para } k = 1,2,\ldots,\left\lfloor \frac{P(n_{sqi})}{2} \right\rfloor. \quad (4.22)$$

Neste trabalho, o alinhamento é realizado no processo C170 (*Alinhamento das CW's*) por sub-quadro. O processo recebe as $CW_{anterior}$ e a CW_{atual}, relativas ao sub-quadro anterior e ao sub-quadro atual, e aplica-se o método da correlação cruzada, mencionado anteriormente neste capítulo, determinando o deslocamento τ, correspondente à máxima correlação. O deslocamento τ permite o alinhamento da CW_{atual} com a $CW_{anterior}$, sendo então incorporado à CW_{atual} utilizando-se uma adição linear da fase aos seus coeficientes de Fourier o que corresponde ao deslocamento circular da CW_{atual} no domínio do tempo. A CW_{atual} já alinhada passa então a compor a superfície $u(n_i,\phi)$ em conjunto com as CW's anteriores, previamente alinhadas, ficando pronta para o próximo processo C190 (*Normalização da Potência das CW's*), conforme o esquema da Figura 4.9.

A Figura 4.18 mostra o diagrama esquemático para a seqüência das CW's alinhadas, $u[n_{sqi}, \phi]$ após a aplicação do processo C190, sobre as CW's mostradas na Figura 4.8.

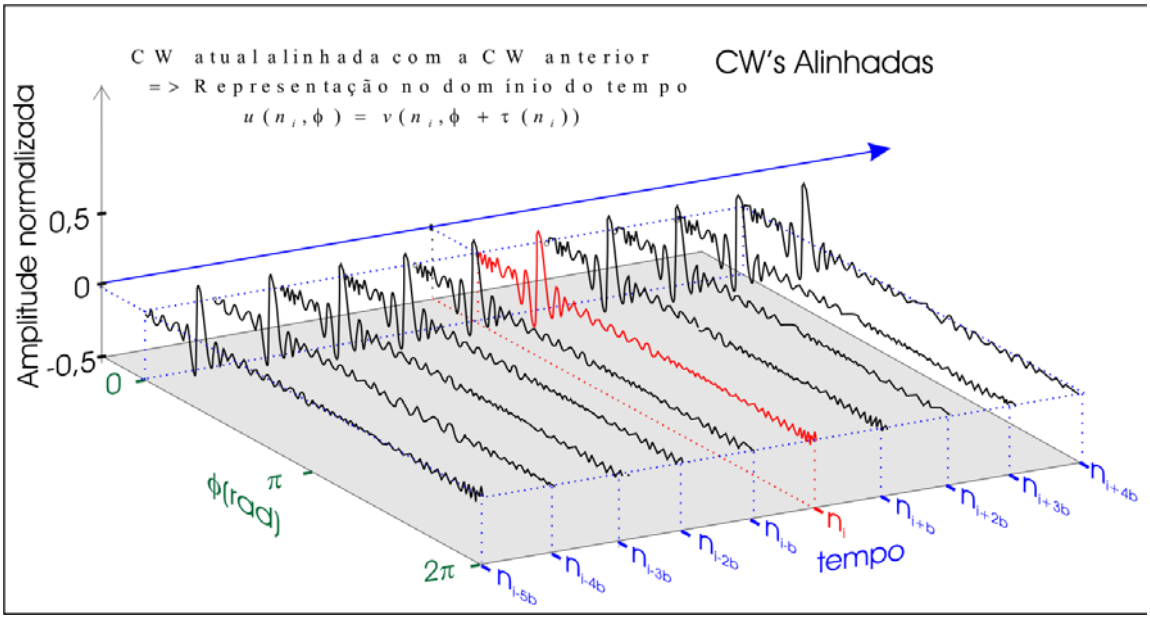

Figura 4.18 – Diagrama esquemático – Representação em duas dimensões das formas de ondas características após a aplicação do processo de alinhamento (processo C170).

Para a aplicação do critério de alinhamento, como mencionado anteriormente neste capítulo, as CW's devem ter a mesma dimensão. Quando as CW's têm dimensões diferentes é necessário aplicar um pré-processamento sobre elas no domínio da freqüência, tais como inserção de zeros, truncamento espectral e acréscimo (ou apêndice) de zeros, de forma que elas tenham a mesma dimensão (ou a mesma quantidade de harmônicas), antes da aplicação do critério de alinhamento. Estes procedimentos permitem a aplicação do mesmo critério de alinhamento, o que facilita a implementação do codificador.

Para facilitar a compreensão e a implementação, o processo C170 (*Alinhamento das CW's*), esquematizado na Figura 4.19, é dividido nos processos C171, C172, C173, C174 e C175 [14], que são descritos a seguir:

Processo C171 (*Inserção de zeros*) – Este processo é utilizado se o pitch tiver múltiplos ou submúltiplos. Na fala natural pode ocorrer duplicação, triplicação ou divisão do pitch ao meio. Assim múltiplos ou submúltiplos do ciclo do pitch podem acontecer, considerando o pitch da CW atual em relação ao pitch da CW anterior. Para prevenir problemas no alinhamento, a CW menor é estendida periodicamente um número inteiro de vezes no processo 171. Este procedimento é realizado para se obter uma maior aproximação do comprimento da CW maior, antes da aplicação do critério de alinhamento. No domínio espectral isto é equivalente a inserir harmônicas com amplitudes com valor igual a zero entre os coeficientes DTFS $\{A_k, B_k\}$ da CW mais curta. No domínio da freqüência, um zero entre duas harmônicas duplica a CW no domínio do tempo, dois zeros, triplica no domínio do tempo, e assim por diante.

Processo C172 (*Truncamento ou acréscimo de zeros*) – Se as CW's tiverem comprimentos diferentes, com a diferença menor que um múltiplo ou submúltiplo do pitch, aplica-se o truncamento ou o acréscimo de zeros no domínio espectral (nos coeficientes DTFS da CW). Executa-se um pré-processamento nas CW's, CW atual e CW anterior para que elas tenham o mesmo comprimento antes da aplicação do critério de alinhamento (Processo C173). O pré-processamento aplicado é:

a) *Truncamento Espectral* - Se o comprimento da CW atual for maior do que o da CW anterior aplica-se o truncamento espectral, desconsiderando-se as harmônicas de maior freqüência, de forma que as CW's fiquem com o mesmo comprimento. O *truncamento espectral* causa uma *contração temporal* de forma que se pode igualar o comprimento de uma CW com a outra de comprimento menor. O truncamento espectral é equivalente a desconsiderar os coeficientes "DTFS" da CW que representam as maiores freqüências, ou seja, desconsidera-se os últimos coeficientes nas seqüências $\{A_k, B_k\}$.
b) *Acréscimo (preenchimento ou apêndice) espectral com zeros* - Se o comprimento da CW atual for menor do que o comprimento da CW anterior completa-se com zeros o final do espectro da CW atual, de forma que elas fiquem com o mesmo comprimento. Este processo acrescenta harmônicas de alta freqüência com amplitudes com valores iguais a zero à CW atual. O acréscimo (ou apêndice) de zeros no domínio espectral causa uma *expansão temporal* de forma que se pode igualar o comprimento de uma CW com a outra CW com comprimento maior. O acréscimo (ou apêndice) de zeros é realizado acrescentando-se zeros às seqüências de coeficientes "DTFS" $\{A_k, B_k\}$ nas posições de maior ordem.

Processo C173 (*Otimização do critério de alinhamento*) – Neste processo calcula-se o deslocamento $\tau(n_{sqi})$ utilizando-se a Equação (4.21) onde $\tau(n_{sqi})$ é a quantidade de deslocamento na fase, necessário para alinhar a CW do sub-quadro atual com a CW do sub-quadro anterior. A subseção 4.3.3.5.1 deste capítulo apresenta o algoritmo desenvolvido para o processo C173 que foi utilizado na implementação deste trabalho.

Processo C174 (*Modificação dos coeficientes DTFS*) – Neste processo incorpora-se o deslocamento $\tau(n_{sqi})$ atualizando os coeficientes DTFS da CW atual utilizando-se a Equação (4.22). A CW atual alinhada, *CW atual alinhada* $\Leftrightarrow$ $u[n_{sqi}, \phi]$ $\Leftrightarrow$ $\{A'_k(n_{sqi}), B'_k(n_{sqi})\}$, fica então disponível para o processo C175 (*Atraso de um sub-quadro*) e, na saída do processo170, é enviada para o próximo processo C190 (*Normalização da Potência das CW's*), conforme os esquemas mostrados nas Figuras 4.9 e 4.19.

Processo C175 (*Atraso por um sub-quadro*) – Este processo envia os coeficientes DTFS *CW atual alinhada* $\Leftrightarrow$ $u[n_{sqi},\phi]$ $\Leftrightarrow$ $\{A'_k(n_{sqi}), B'_k(n_{sqi})\}$, disponíveis na saída do processo C170, para atualizar, na entrada do processo C170, a CW anterior. Neste mesmo instante a CW atual, também na entrada do processo C170, recebe os coeficientes DTFS da CW do próximo sub-quadro a ser alinhado. Assim, as CW's atualizadas, na entrada do processo C170, representadas por:

$$\begin{aligned} &CW\ anterior \Leftrightarrow u[n_{sq(i-1)},\phi] \Leftrightarrow \{A'_k(n_{sq(i-1)}), B'_k(n_{sq(i-1)})\} \Leftarrow \{A'_k(n_{sqi}), B'_k(n_{sqi})\} \\ &CW\ atual \quad \Leftrightarrow v[n_{sqi},\phi] \quad \Leftrightarrow \{A_k(n_{sqi}), B_k(n_{sqi})\} \qquad \Leftarrow \{A_k(n_{sq(i+1)}), B_k(n_{sq(i+1)})\} \end{aligned}$$

estão disponíveis para o próximo processo de alinhamento, onde $n_{sq(i-1)}, n_{sqi}, n_{sq(i+1)}$ correspondem respectivamente às posições das CW's *anterior*, *atual e posterior*.

A Figura 4.19 mostra o diagrama de blocos do processo C170 (*Alinhamento das CW's*) que foi implementado neste trabalho.

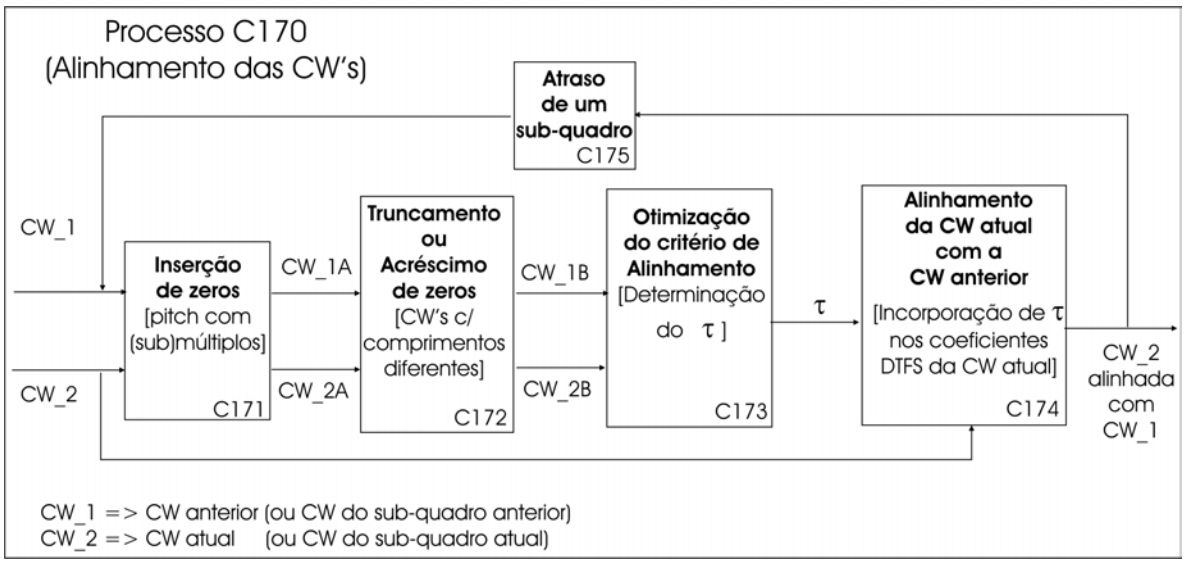

Figura 4.19 – Diagrama de blocos do processo C170 – *Alinhamento das CW's*.

O objetivo do processo C170 é realizar o alinhamento da CW atual (sub-quadro atual) com a CW anterior (sub-quadro anterior) no domínio da DTFS utilizando-se a correlação cruzada entre os coeficientes das duas CW's. Esta operação coloca as CW's em fase, ou seja, alinham no tempo as principais características das CW's. O processo permite realizar um alinhamento fracionário, onde o deslocamento T pode assumir qualquer valor real na faixa, $0 \le T < P(n_{sqi})$, ou seja, com uma resolução no tempo menor ou maior do que o período de amostragem T_s, sem a necessidade de alteração na freqüência de amostragem do sinal residual $e[n]$, onde T e $\tau(n_{sqi})$, deslocamentos de fase, estão relacionados por:

$$\tau(n_{sqi}) = \frac{2\pi T}{P(n_{sqi})}, \tag{4.23}$$

onde $\tau(n_{sqi})$ também pode assumir qualquer valor real na faixa $0 \le \tau(n_{sqi}) < 2\pi$. Na implementação o deslocamento $\tau(n_{sqi})$ foi calculado com uma resolução no tempo de até $\frac{1}{4}T_s$.

Procedimentos para a implementação do processo C170: Na implementação, o processo C170 (*Alinhamento das CW's*) é executado em um dos três módulos (módulo 1, módulo 2 ou módulo 3). Na escolha do módulo é utilizado o parâmetro C, definido por:

$$C = \frac{P(n_{sqi})}{P(n_{sq(i-1)})} \tag{4.24}$$

para $P(n_{sqi}) \ge P(n_{sq(i-1)})$ (múltiplos), ou:

$$C = \frac{P(n_{sq(i-1)})}{P(n_{sqi})} \tag{4.25}$$

para $P(n_{sqi}) \le P(n_{sq(i-1)})$ (submúltiplos). Assim os módulos são definidos como:

a) **<u>Módulo 1</u>**: As CW's tem a mesma dimensão (C=1). Este módulo envolve os processos C173 e C174:
   - Processo 173 -> Equação (4.21) (Cálculo de $\tau(n_{sqi})$);
   - Processo 174 -> Equação (4.22) (Modificação dos coeficientes da DTFS).

b) **<u>Módulo 2</u>**: As CW's tem dimensões diferentes (1 < C < 2) (sem múltiplos ou sub-múltiplos do pitch). Envolve os processos C172, C173 e C174:
   - Processo 172 -> Truncamento ou acréscimo de zeros no domínio espectral (coeficientes da DTFS);
   - Processo 173 -> Equação (4.21) (Cálculo de $\tau(n_{sqi})$);
   - Processo 174 -> Equação (4.22) (Modificação dos coeficientes da DTFS).

c) **<u>Módulo 3</u>**: As CW's tem dimensões diferentes e ocorrem múltiplos ou submúltiplos do pitch ($C \ge 2$). Envolve os processos C171, C172, C173 e C174:
   - Processo 171 => Inserção de zeros (Pitch com múltiplos ou submúltiplos).

- Processo 172 -> Truncamento ou acréscimo de zeros no domínio espectral (coeficientes da DTFS)
- Processo 173 -> Equação (4.21) (Cálculo de $\tau(n_{sqi})$);
- Processo 174 -> Equação (4.22) (Modificação dos coeficientes DTFS).

4.3.3.5.1 Algoritmo para o Processo C173 (*Otimização do critério de alinhamento* das CW's)

(a) Proposta Geral do Algoritmo

O algoritmo tem como objetivo determinar o valor de $\tau(n_{sqi})$ que maximiza a correlação cruzada entre duas CW's (atual e anterior) através da aplicação do critério de alinhamento obtido pela Equação (4.21) onde $\tau(n_{sqi})$ é o valor do deslocamento de fase, para a CW_atual em relação à CW_anterior, escolhido entre os valores assumidos por $\tau'(n_{sqi})$ na faixa de $0 \le \tau'(n_{sqi}) < 2\pi$.

O algoritmo é dividido em quatro fases com a intenção de reduzir a quantidade de cálculos e de processamento. As fases são: FASE B, FASE M, FASE A e a FASE FINAL, conforme mostra a Figura 4.20.

FASE B *(Pesquisa do $\tau(n_{sqi})$ em baixa resolução)*: Durante a FASE B, o cálculo da correlação é realizado para valores de τ' mais espaçados com um valor definido por $\Delta\tau_B$, na faixa $0 \le \tau'(n_{sqi}) < 2\pi$. O objetivo é identificar duas regiões onde ocorrem os maiores valores para a correlação cruzada, para que nas fases seguintes, seja realizada uma pesquisa mais apurada para o deslocamento de fase. Assim uma constante Div_B (relacionada com $\Delta\tau_B$) é definida no início do algoritmo. Esta constante impõe uma pesquisa com *baixa resolução* (B) para o deslocamento de fase, na procura por dois deslocamentos de fase (τ_{1B} e τ_{2B}) com as maiores correlações cruzadas.

FASE M *(Pesquisa do $\tau(n_{sqi})$ em resolução média)*: Durante a FASE M, o cálculo da correlação é realizado para os valores de τ' espaçados com um valor definido por $\Delta\tau_M$, onde $\Delta\tau_M < \Delta\tau_B$, nas faixas $\tau_{1B} - \frac{Faixa\_M}{2} \le \tau'(n_{sqi}) \le \tau_{1B} + \frac{Faixa\_M}{2}$ e $\tau_{2B} - \frac{Faixa\_M}{2} \le \tau'(n_{sqi}) \le \tau_{2B} + \frac{Faixa\_M}{2}$. Assim aumenta-se a resolução e diminui-se a faixa de procura em relação à fase anterior, FASE B. O valor da constante Faixa_M, que é obtido em número de $\Delta\tau_M$, é a faixa (ou intervalo) de valores onde τ' é pesquisado em torno de τ_{1B} e τ_{2B}. A Faixa_M e Div_M são os parâmetros, apresentados no início do algoritmo,

que definem a pesquisa com a *resolução média* (M) para o deslocamento de fase, na procura por dois deslocamentos de fase (τ_{1M} e τ_{2M}) com as maiores correlações cruzadas.

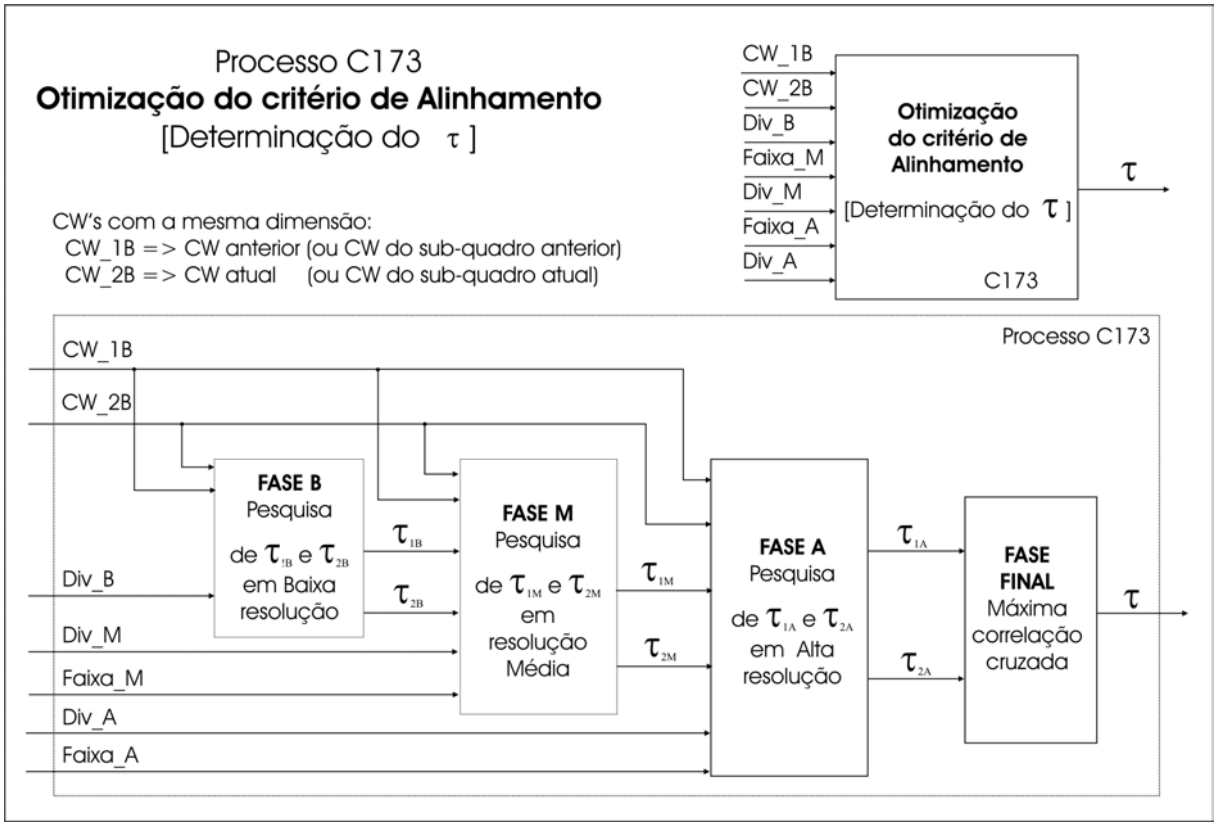

Figura 4.20 – Diagrama de blocos do processo C173 – Otimização do critério de *Alinhamento (Determinação do* $\tau(n_{sqi})$.

FASE A *(Pesquisa do* $\tau(n_{sqi})$ *em alta resolução)*: Durante a FASE A, o cálculo da correlação é realizado para os valores de τ' espaçados com um valor definido por $\Delta\tau_A$, onde $\Delta\tau_A < \Delta\tau_M < \Delta\tau_B$, nos intervalos $\tau_{1M} - \frac{Faixa\_A}{2} \le \tau'(n_{sqi}) \le \tau_{1M} + \frac{Faixa\_A}{2}$ e $\tau_{2M} - \frac{Faixa\_A}{2} \le \tau'(n_{sqi}) < \tau_{2M} + \frac{Faixa\_A}{2}$. Assim aumenta-se a resolução e diminui-se a faixa de procura em relação à FASE M. O valor da constante Faixa_A, que é obtido em número de $\Delta\tau_A$, é a faixa (ou intervalo) de valores de τ' que é pesquisado em torno de τ_{1M} e τ_{2M}. A Faixa_A e Div_A são parâmetros, apresentados no início do algoritmo, que definem a pesquisa com *resolução alta* (A) para o deslocamento de fase, na procura por dois deslocamentos de fase (τ_{1A} e τ_{2A}) com as maiores correlações cruzadas.

FASE_FINAL (*Escolha final para o valor de* τ**):** A partir dos valores das correlações cruzadas calculadas em correspondência a τ_{1A} e τ_{2A} escolhe-se para o valor de τ um dos argumentos (τ_{1A} ou τ_{2A}) que tiver a maior correlação cruzada. A Figura 4.21 mostra em detalhes as fases do algoritmo.

(b) Equações utilizadas no algoritmo:

(b1) Na determinação do $\tau(n_{sqi})$ é utilizada a Equação (4.21) reescrita aqui por conveniência:

$$\tau(n_{sqi}) = \arg\underset{0\le\tau'<2\pi}{máx}\sum_{k=1}^{M}\{[A_k(n_{sq(i-1)})A_k(n_{sqi}) + B_k(n_{sq(i-1)})B_k(n_{sqi})]\cos(k\tau') + [B_k(n_{sq(i-1)})A_k(n_{sqi}) - B_k(n_{sqi})A_k(n_{sq(i-1)})]\text{sen}(k\tau')\}$$

Para a utilização no algoritmo a Equação (4.21) pode ser escrita como:

$$\tau(n_{sqi}) = \arg\underset{0\le\tau'<2\pi}{máx} f(\tau') \tag{4.26}$$

onde:

$$f(\tau') = \sum_{k=1}^{M}\{[A_k(n_{sq(i-1)})A_k(n_{sqi}) + B_k(n_{sq(i-1)})B_k(n_{sqi})]\cos(k\tau') + [B_k(n_{sq(i-1)})A_k(n_{sqi}) - B_k(n_{sqi})A_k(n_{sq(i-1)})]\text{sen}(k\tau')\} \tag{4.27}$$

representa a correlação cruzada entre as CW's expressa em termos dos coeficientes da DTFS.

(b2) A Equação (4.23) $\tau(n_{sqi}) = \dfrac{2\pi T}{P(n_{sqi})}$ é utilizada na definição da resolução angular (fase) $\Delta\tau$.

(c) Entrada do Algoritmo (Dados disponíveis no início):

O algoritmo recebe:

- As CW's anterior e atual (representadas pelos coeficientes da DTFS) ⇔ CW_1B e a CW_2B (têm nesta posição a mesma dimensão), onde:

$$CW\_1B\ (CW\_anterior - Alinhada) \Leftrightarrow u[n_{sq(i-1)}, \phi] \Leftrightarrow \{A'_k(n_{sq(i-1)}), B'_k(n_{sq(i-1)})\}$$
$$CW\_2B\ (CW\_atual - a\ ser\ Alinhada) \Leftrightarrow v[n_{sqi}, \phi] \Leftrightarrow \{A_k(n_{sqi}), B_k(n_{sqi})\}$$

- As informações sobre a resolução e a faixa de pesquisa de τ' em cada FASE:
  FASE B: - **Div_B** -> Fator para a determinação da resolução $\Delta\tau_B$.

Obs.: Nesta fase a faixa de procura de $\tau(n_{sqi})$ é padronizada em $0 \le \tau' < 2\pi$.

FASE M: - **Div_M** -> Fator para a determinação da resolução $\Delta\tau_M$.
- **Faixa_M** -> Faixa de variação de τ' em número de $\Delta\tau_M$.

FASE A: - **Div_A** -> Fator para a determinação da resolução $\Delta\tau_A$.
- **Faixa_A** -> Faixa de variação de τ' em número de $\Delta\tau_A$

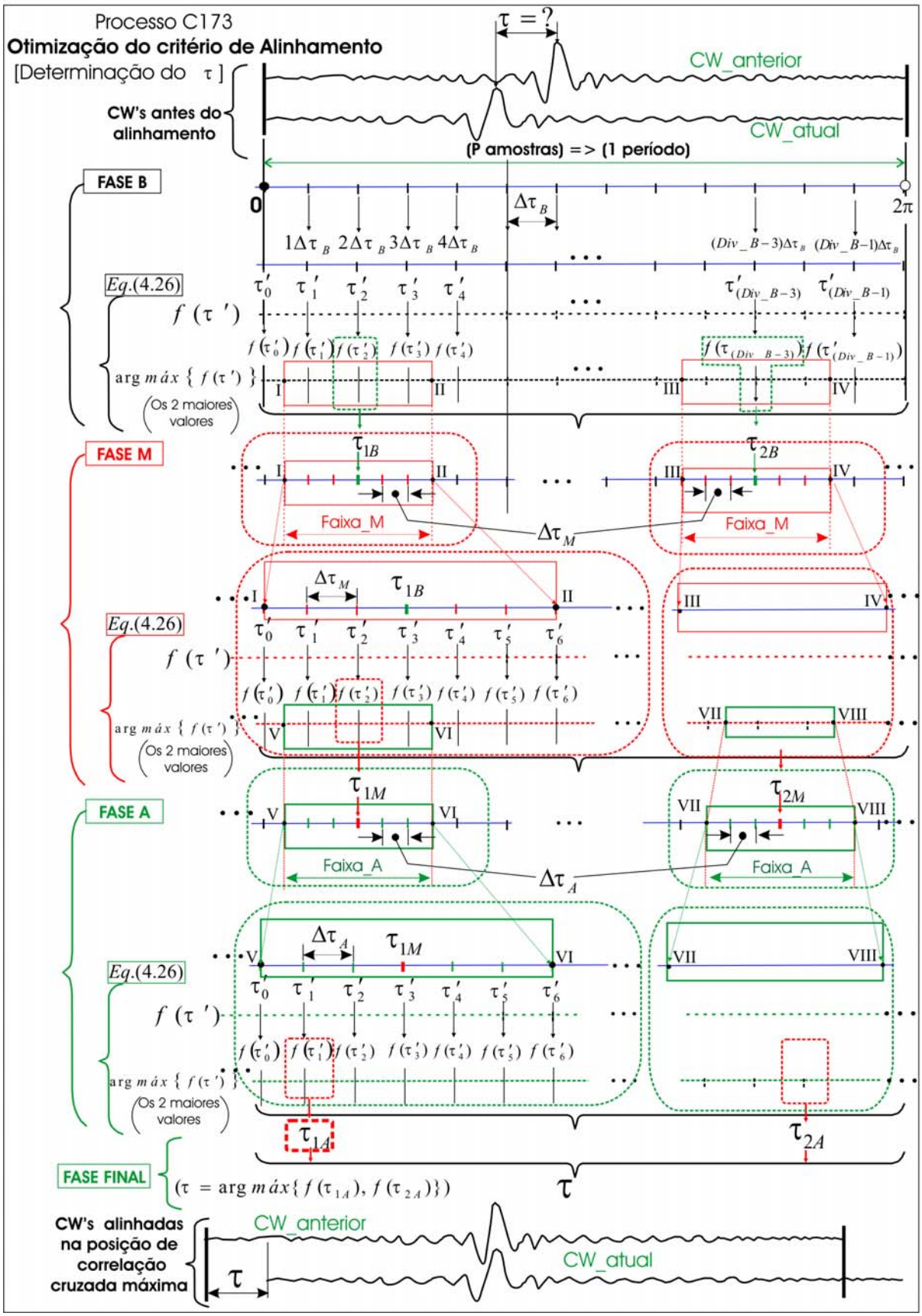

Figura 4.21 – Diagrama esquemático- Detalhes do processo C173 – Otimização do critério de Alinhamento.

(d) Constantes Calculadas no Algoritmo:

- $\Delta\tau_B$ - Resolução de fase entre os valores assumidos por τ' na FASE B;
- $\Delta\tau_M$ - Resolução de fase entre os valores assumidos por τ' na FASE M;
- $\Delta\tau_A$ - Resolução de fase entre os valores assumidos por τ' na FASE A;

As resoluções de fase são obtidas por

$$\Delta\tau_B = \frac{2\pi}{Div\_B * P(n_{sqi})}, \quad \Delta\tau_M = \frac{2\pi}{Div\_M * P(n_{sqi})} \quad \text{e} \quad \Delta\tau_A = \frac{2\pi}{Div\_A * P(n_{sqi})}.$$

Utilizando a Equação (4.23), $\tau(n_{sqi}) = \frac{2\pi T}{P(n_{sqi})}$, pode-se verificar a relação entre a resolução de fase $\Delta\tau$ e a resolução temporal ΔT. Assim a Equação (4.23) torna-se

$$T = \frac{\tau(n_{sqi})P(n_{sqi})}{2\pi} \quad => \quad \Delta T = \frac{\Delta\tau(n_{sqi})P(n_{sqi})}{2\pi}$$

$$\Delta T = \frac{\frac{2\pi}{Div * P(n_{sqi})}P(n_{sqi})}{2\pi} => \Delta T = \frac{1}{Div} \text{ (amostras) que corresponde a } \Delta T = \frac{1}{Div}T_S$$

onde ΔT é a resolução temporal e T_S é o período de amostragem do sinal residual. Assim as resoluções temporais em função do período de amostragem tornam-se:

$$\Delta T_B = \frac{1}{Div\_B}T_S, \quad \Delta T_M = \frac{1}{Div\_M}T_S \quad \text{e} \quad \Delta T_A = \frac{1}{Div\_A}T_S. \tag{4.28}$$

Neste trabalho foram utilizados os valores Div_B = 1,0, Div_M = 2,0 e Div_A = 4,0 que correspondem respectivamente às resoluções temporais $\Delta T_B = T_S$, $\Delta T_M = \frac{1}{2}T_S$ e $\Delta T_A = \frac{1}{4}T_S$.

(e) Saída do Algoritmo (por FASE)

FASE B:

- τ_{1B} e τ_{2B} (os dois deslocamentos de fase onde a correlação cruzada é maior na FASE B);

FASE M:

- τ_{1M} e τ_{2M} (os dois deslocamentos onde a correlação cruzada é maior na FASE M);

FASE F:

- τ_{1A} e τ_{2A} (os dois deslocamentos onde a correlação cruzada é maior na FASE A);

FASE FINAL:

- τ (Valor final para o deslocamento de fase com a correlação cruzada máxima).

(f) Procedimentos (O algoritmo):

Passo 1:

(1) Recebe as **CW anterior** e a **CW atual** (coeficientes da DTFS) – ver item (c);

(2) **CW anterior** e a **CW atual** tornam-se disponíveis para os próximos passos.

Passo 2:

(1) Recebe os parâmetros que definem as resoluções e as faixas (ou intervalos) de pesquisa de τ':

Div_B, Div_M, Div_A, Faixa_M e Faixa_A;

(2) Estes parâmetros tornam-se disponíveis para os próximos passos.

Passo 3:

(1) Calcula as resoluções para a pesquisa do deslocamento de fase em cada fase: $\Delta\tau_B$, $\Delta\tau_M$ e $\Delta\tau_A$ utilizando-se as Equações (4.28);

(2) As resoluções tornam-se disponíveis para as próximas fases.

Passo 4: FASE B – Pesquisa do deslocamento de fase em Baixa Resolução
(Determina τ_{1B} e τ_{2B})

(1) Calcula os valores de τ' usando a resolução $\Delta\tau_B$ na faixa de $0 \le \tau' < 2\pi$;

(2) Calcula $f(\tau')$ utilizando-se a Equação (4.27);

(3) Aplica a Equação (4.26) determinando τ_{1B} e τ_{2B} que correspondem aos dois maiores valores para $f(\tau')$;

(4) τ_{1B} e τ_{2B} tornam-se disponíveis para a próxima fase, a FASE M.

Passo 5: **FASE M - Pesquisa do deslocamento de fase em Média Resolução**
(Determina τ_{1M} e τ_{2M})

(1) Aplica-se a Faixa_M (usando a resolução $\Delta\tau_M$) em torno de τ_{1B} e τ_{2B}, calculando os valores de τ'. Para τ_{1B} vale o intervalo $\tau_{1B} - \frac{Faixa\_M}{2} \le \tau'(n_{sqi}) \le \tau_{1B} + \frac{Faixa\_M}{2}$ que na Figura 4.21 corresponde a $\tau_I \le \tau'(n_{sqi}) \le \tau_{II}$;

Para τ_{2B} vale o intervalo $\tau_{2B} - \frac{Faixa\_M}{2} \le \tau'(n_{sqi}) \le \tau_{2B} + \frac{Faixa\_M}{2}$ que na Figura 4.21 corresponde a $\tau_{III} \le \tau'(n_{sqi}) \le \tau_{IV}$;

(2) Para o intervalo $\tau_I \le \tau'(n_{sqi}) \le \tau_{II}$, em torno de τ_{1B}, calcula-se $f(\tau')$ utilizando-se a Equação (4.27); e, aplica-se a Equação (4.26) determinando τ_{1M} que corresponde ao valor máximo para $f(\tau')$;

(3) Para o intervalo $\tau_{III} \le \tau'(n_{sqi}) \le \tau_{IV}$ em torno de τ_{2B}, calcula-se $f(\tau')$ através da Equação (4.27); e, aplica-se a Equação (4.26) e determinando-se τ_{2M} que corresponde ao valor máximo para $f(\tau')$;

(4) τ_{1M} e τ_{2M} tornam-se disponíveis para a próxima fase, a FASE A.

Passo 6: **FASE A - Pesquisa do deslocamento de fase em Alta Resolução**
(Determina τ_{1A} e τ_{2A})

(1) Aplica-se a Faixa_A (usando a resolução $\Delta\tau_A$) em torno de τ_{1M} e τ_{2M}, calculando-se os valores de τ'.

(2) Para τ_{1M} vale o intervalo $\tau_{1M} - \frac{Faixa\_A}{2} \le \tau'(n_{sqi}) \le \tau_{1M} + \frac{Faixa\_A}{2}$ que na Figura 4.21 corresponde a $\tau_V \le \tau'(n_{sqi}) \le \tau_{VI}$;

Para τ_{2M} vale o intervalo $\tau_{2M} - \frac{Faixa\_A}{2} \le \tau'(n_{sqi}) \le \tau_{2M} + \frac{Faixa\_A}{2}$ que na Figura 4.21 corresponde a $\tau_{VII} \le \tau'(n_{sqi}) \le \tau_{VIII}$;

(3) Para o intervalo $\tau_V \le \tau'(n_{sqi}) \le \tau_{VI}$, em torno de τ_{1M}, calcula-se $f(\tau')$ utilizando-se Equação (4.27); e, aplica-se a Equação (4.26) determinando τ_{1A} que corresponde ao valor máximo para $f(\tau')$;

(4) Para o intervalo $\tau_{VII} \le \tau'(n_{sqi}) \le \tau_{VIII}$ em torno de τ_{2M}, calcula-se $f(\tau')$ utilizando-se a Equação (4.27); e, aplica-se a Equação (4.26) determinando τ_{2A} que corresponde ao valor máximo para $f(\tau')$;

(5) τ_{1A}, τ_{2A}, $f(\tau_{1A})$ e $f(\tau_{2A})$ tornam-se disponíveis para a próxima fase, a FASE FINAL.

Passo 7: **FASE FINAL** (Determina τ)

(1) Compara $f(\tau_{1A})$ e $f(\tau_{2A})$, o maior valor define τ, ou seja, se $f(\tau_{1A}) > f(\tau_{2A})$ então $\tau = \tau_{1A}$, caso contrário $\tau = \tau_{2A}$.

(2) $\tau(n_{sqi}) = \tau$ torna-se disponível para o próximo processo, Processo C174 – Alinhamento da CW atual com a CW anterior (Incorporação de τ aos coeficientes da DTFS).

(3) Fim.

4.3.3.6 Cálculo da Potência e Normalização das CW's

As CW's alinhadas no processo C170 tornam-se disponíveis para serem normalizadas para terem potência média unitária através do processo C190 *(Normalização da Potência das CW's)*. Antes do processo C190, as CW's passam pelo processo C180 *(Cálculo da Potência das CW's)* onde a potência média de cada CW é calculada. As CW's são normalizadas em potência para separar a potência e a forma das CW's para que elas possam ser quantizadas separadamente para permitir uma codificação mais eficiente. A potência média de uma CW extraída em n_{sqi}, $Pot(n_{sqi})$ pode ser definida como:

$$Pot(n_{sqi}) = \frac{1}{2\pi}\int_0^{2\pi}\left|u[n_{sqi},\phi]\right|^2 d\phi\,. \tag{4.29}$$

Em função dos coeficientes de Fourier pode ser expressa por [14]:

$$Pot(n_{sqi}) = \sum_{k=1}^{\lfloor P(n_{sqi})/2 \rfloor}\left(\frac{1}{\chi_k}\left(\left(A'_k(n_{sqi})\right)^2 + \left(B'_k(n_{sqi})\right)^2\right)\right) \tag{4.30}$$

onde χ_k é um fator apresentado na Equação (4.3). A CW normalizada em potência pode ser representada por:

$$u_N(n_{sqi},\phi) = \sum_{k=1}^{\lfloor P(n_{sqi})/2 \rfloor}\left(\left(A_{Nk}(n_{sqi})\cos(k\phi) + B_{Nk}(n_{sqi})\,\mathrm{sen}(k\phi)\right)\right) \tag{4.31}$$

onde:

$$A_{Nk}(n_{sqi}) = \frac{A'_k(n_{sqi})}{\sqrt{Pot(n_{sqi})}}$$

$$B_{Nk}(n_{sqi}) = \frac{B'_k(n_{sqi})}{\sqrt{Pot(n_{sqi})}}\,. \tag{4.32}$$

No processo C180 a potência das CW's previamente alinhadas $u(n_{sqi},\phi) \Leftrightarrow \left\{A'_k(n_{sqi}), B'_k(n_{sqi})\right\}$ são calculadas utilizando-se a Equação (4.30) a partir dos seus coeficientes de Fourier $\left\{A'_k(n_{sqi}), B'_k(n_{sqi})\right\}$. No processo C190 as CW's são normalizadas em potência dividindo cada coeficiente de Fourier por $\sqrt{Pot(n_{sqi})}$, as Equações (4.32), passando a serem representadas pela Equação (4.31).

As CW's (normalizadas em potência $u_N(n_{sqi},\phi) \Leftrightarrow \{A_{Nk}(n_{sqi}), B_{Nk}(n_{sqi})\}$), utilizando-se os seus coeficientes de Fourier $\{A_{Nk}(n_{sqi}), A_{Nk}(n_{sqi})\}$ e a potência das CW's, $Pot(n_{sqi})$, são parâmetros que tornam-se disponíveis para a compressão (codificação) e/ou transmissão para o decodificador. A Figura 4.22 mostra o diagrama esquemático para a seqüência de CW's normalizadas após a aplicação do processo C190 sobre as CW's representadas na Figura 4.18.

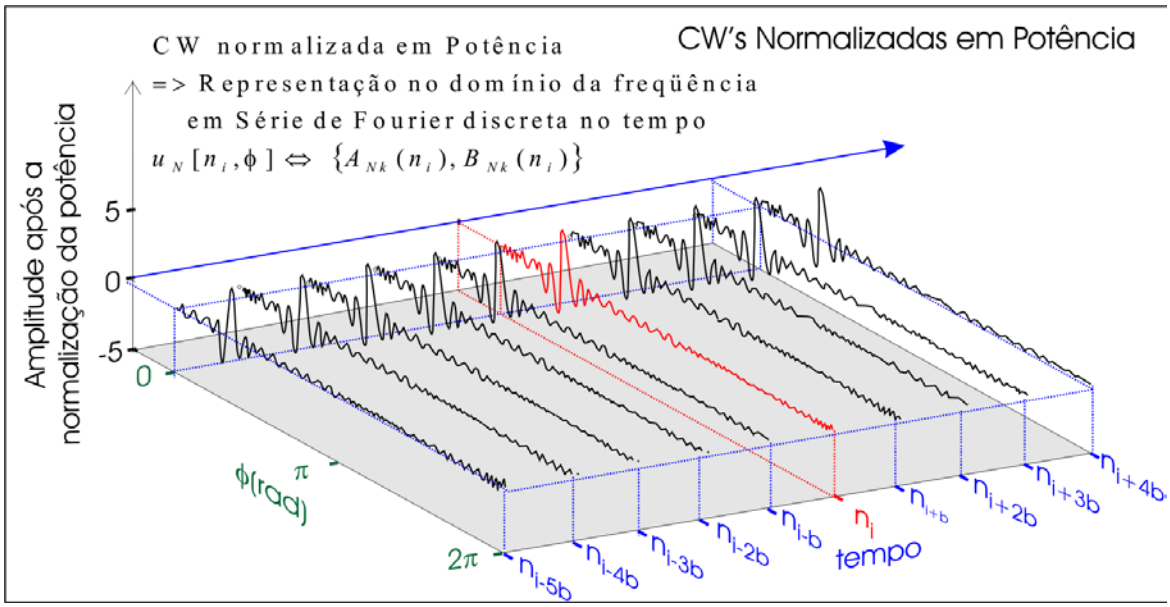

Figura 4.22 – Diagrama esquemático – Representação em duas dimensões das formas de ondas características após a aplicação do processo de Normalização da Potência das CW's (Processo C190).

Considerações sobre a normalização da potência [14]

- O processo de normalização pode ser colocado antes ou após o processo de alinhamento. A ordem dos cálculos não afeta a representação bidimensional $u_N(n_{sqi},\phi)$ (normalizada e alinhada).
- Após a normalização cada CW tem uma potência média de uma unidade.

4.3.3.7 Parâmetros de Saída do Processo C100 (Bloco de Análise)

O processo C100 (Bloco de Análise) apresenta na sua saída os seguintes parâmetros que se tornam disponíveis para a compressão (no processo C300 – Codificação dos Parâmetros) ou para a transmissão ao decodificador (no processo D200 – Sintetizador):

(1) Os coeficientes LSFs $\{a_k\}_{LSF(l)}$ ou $\{a_k\}_{LSF(qd\_Atual)}$ =>>(1 vetor(Np ordenadas)/quadro);

(2) O pitch $P(n_l)$ ou $P(qd\_Atual)$ ==========>> (1 escalar/quadro);

(3) A potência das CW's ($Pot(n_{sqi})$) ou $Pot(sq\_Atual)$ =====>> (8 escalares/quadro);

(4) As CW's $u_N(n_{sqi},\phi) \Leftrightarrow \{A_{Nk}(n_{sqi}), B_{Nk}(n_{sqi})\}$ ==========>> (8 conjuntos/quadro).

Neste trabalho, durante a implementação os parâmetros, sem compressão, são enviados ao decodificador, processo D200 (Bloco de Síntese), através de um arquivo "par_wi.c" gravado com dados binários com a seguinte formatação:

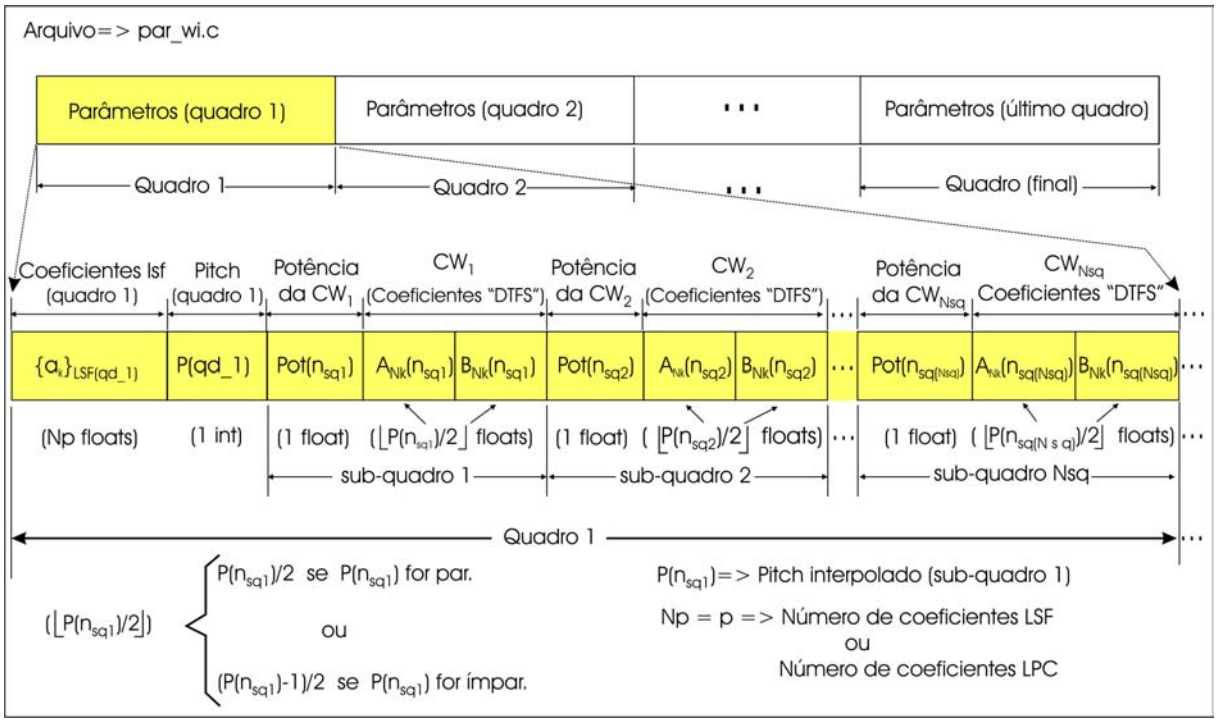

Figura 4.23 – Diagrama esquemático – Representação do arquivo "par_wi.c" utilizado na implementação do codificador WI (processo C100 – Bloco de Análise) para a transmissão dos parâmetros, sem compressão, ao decodificador (processo D200 – Bloco de Síntese).

4.3.3.8 Resumo do Processo C100 (Bloco de Análise)

O estágio de análise apresentou a seguinte seqüência de procedimentos resumidos no item I e descritos de forma sucinta nos itens II, III, IV, V, VI e VII.

(I) O Diagrama de Blocos com Indicação dos Parâmetros

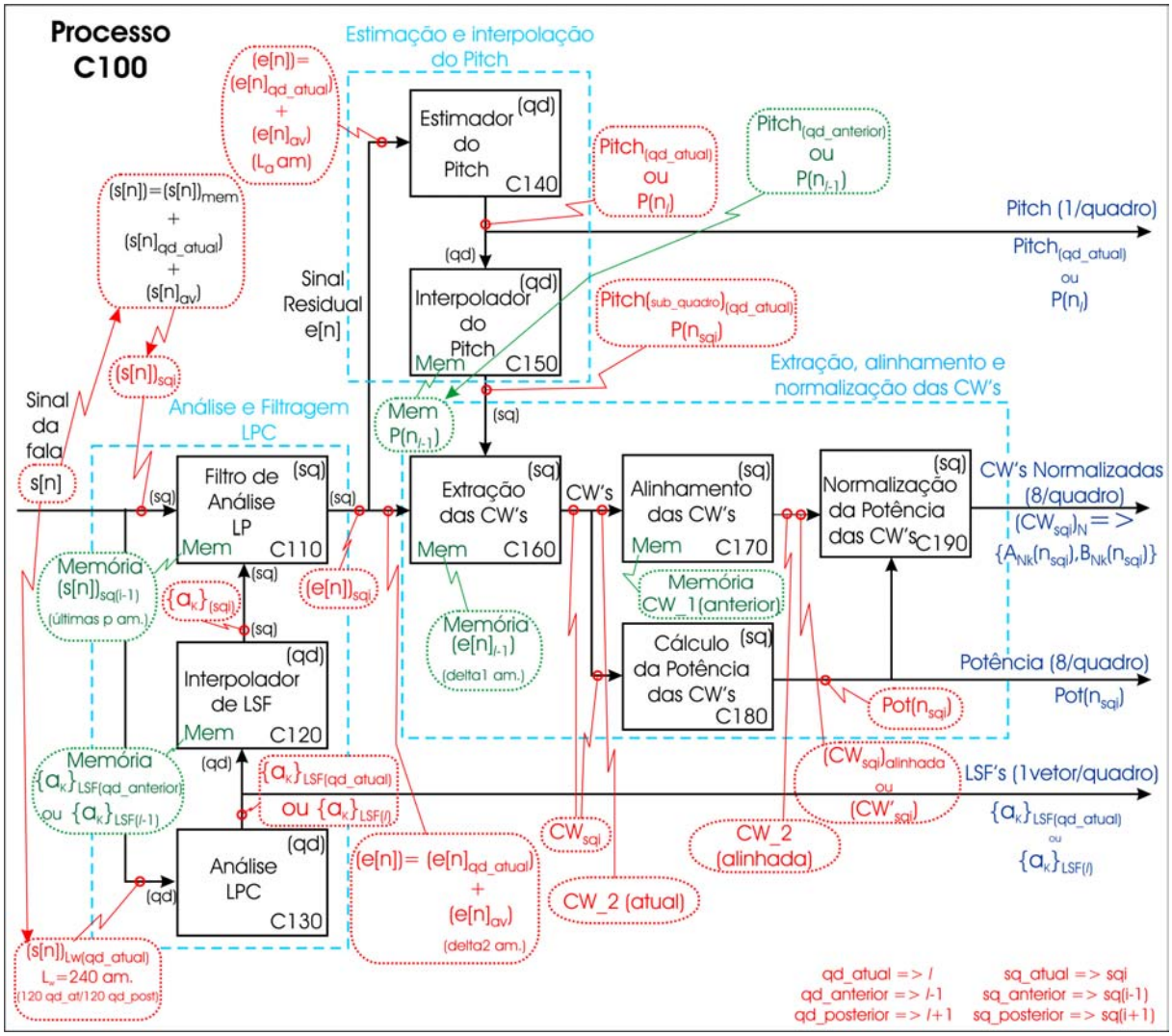

Figura 4.24 – Diagrama de blocos do bloco de análise expandido (processo C100) do codificador WI indicando os parâmetros de entrada e de saída nos blocos.

(II) A Preparação do sinal

1. Sinal de entrada $s[n]$ => sinal digital da fala (original)
   Freqüência de amostragem $fs = 11025\ Hz$.
2. Alguns processos são realizados uma vez por quadro enquanto outros são realizados uma vez por sub-quadro.
   (a) Processamentos por quadros => Divisão de $s[n]$ em quadros consecutivos:
   $\ldots, (s[n]_1), (s[n]_2) \ldots, (s[n]_{l-1}), (s[n]_l), (s[n]_{l+1}) \ldots$ onde $(s[n]_l)$ representa o quadro atual, $(s[n]_{l-1})$ o quadro anterior (ou passado) e $(s[n]_{l+1})$ o quadro posterior (ou em avanço).
   Comprimento do quadro => L_q = 160 amostras (14,5 ms).
   (b) Processamentos por sub-quadros => Divisão de $(s[n]_l)$ em sub-quadros consecutivos:

$(s[n])_{sq0}, (s[n])_{sq1}, \ldots, (s[n])_{sq(i-1)}, (s[n])_{sqi}, (s[n])_{sq(i+1)}, \ldots, (s[n])_{sq(Nsq-1)}$ onde $(s[n])_{sqi}$ representa o sub-quadro atual, $(s[n])_{sq(i-1)}$ o sub-quadro anterior (ou passado) e $(s[n])_{sq(i+1)}$ o sub-quadro posterior (ou em avanço).

Comprimento do sub-quadro => L_{sq} = 20 amostras (1,8 ms).

Na implementação do trabalho foi usado Nsq = 8 e portanto $(s[n])_{sq0}, (s[n])_{sq1}, \ldots, (s[n])_{sq7}$.

(III) A Análise com Predição Linear (Análise LP)

Os processos envolvidos na análise com predição linear (Análise LP):

Blocos C110, C120 e C130.

Janela de análise:

- Comprimento L_W = 240 amostras (120 amostras - quadro atual/ 120 amostras - quadro posterior).
- Posição de L_W => (centro da janela na fronteira entre o quadro atual e o quadro posterior).

Bloco C130 (Análise LP):

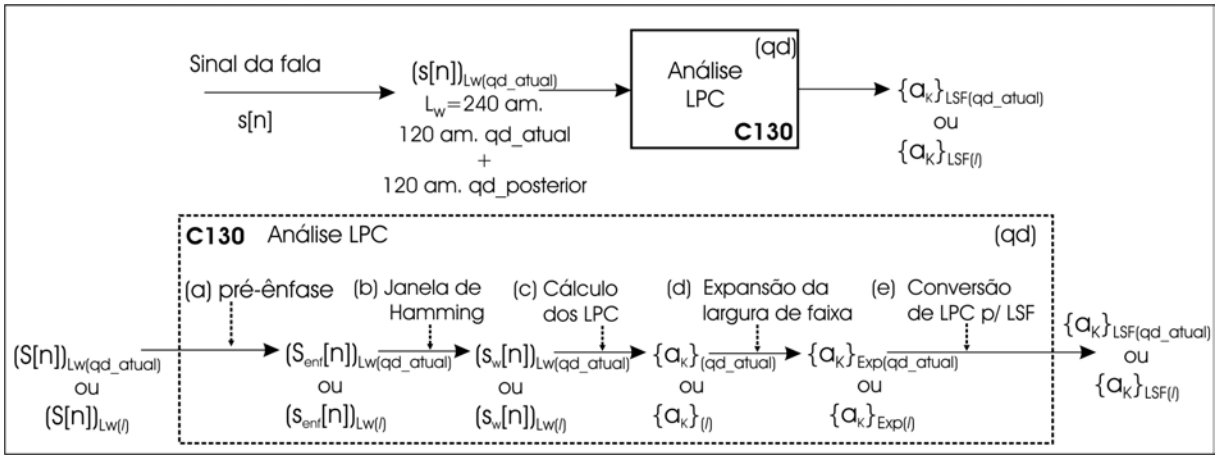

Figura 4.25 – Diagrama de blocos do (processo C130 – Análise LPC) componente do (processo C100 – bloco de Análise) do codificador WI com indicação dos processamentos e dos parâmetros de entrada e de saída.

Entrada: Sinal $(s[n])_{Lw(qd\_atual)}$ na janela de análise.

Saída: O vetor de coeficientes LSF do quadro atual $\{a_k\}_{LSF(qd\_atual)}$.

Processamento:

Pré-ênfase => Equação (2.76);

Janela de Hamming => Equação (2.21);

Cálculo dos vetores LPC => procedimentos descritos na seção 2.5.1.2;

Expansão da largura de faixa => procedimentos descritos na seção 2.5.2;

Conversão dos LPC para os LSF => procedimentos descritos na seção 2.5.1.3.1.

Notação dos parâmetros:

Coeficientes LPC

$\{a_k\}_{(qd\_atual)} = (a_1, a_2, \cdots, a_p) \quad onde \quad 1 \le k \le p$.

Expansão da largura de faixa

$\{a_k\}_{Exp(qd\_atual)} = (a_{1Exp}, a_{2Exp}, \cdots, a_{pExp}) \quad onde \quad 1 \le k \le p$.

Coeficientes LSF

$\{a_k\}_{LSF(qd\_atual)} = (a_{1LSF}, a_{2LSF}, \cdots, a_{pLSF}) \quad onde \quad 1 \le k \le p$.

Bloco C120 (Interpolador de LSF):

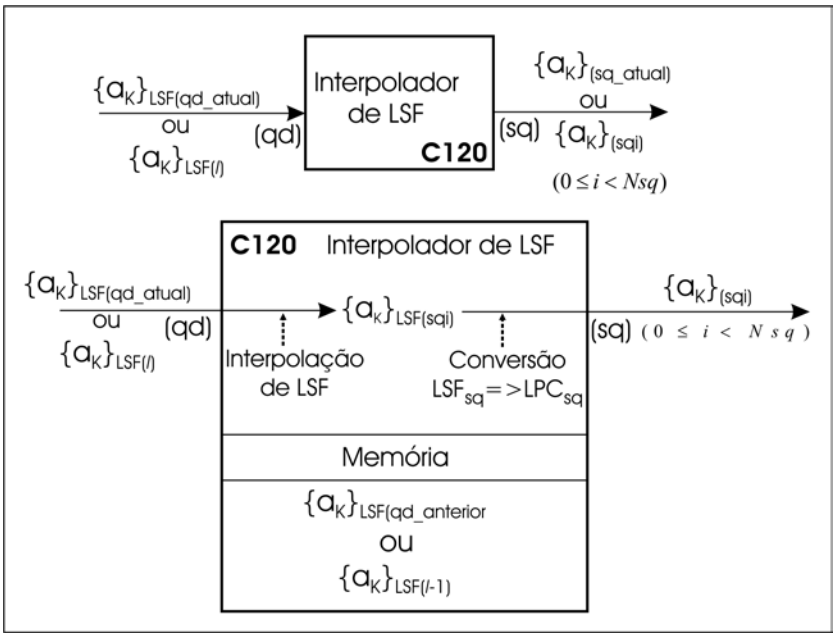

Figura 4.26 – Diagrama de blocos do (processo C120 – Interpolador de LSF) componente do (processo C100 – bloco de Análise) do codificador WI com indicação dos processamentos e dos parâmetros de entrada e de saída.

Entrada: O vetor de coeficientes LSF do quadro atual => $\{a_k\}_{LSF(qd\_atual)}$ ou $\{a_k\}_{LSF(l)}$.

Na memória: O vetor de coeficientes LSF do quadro anterior => $\{a_k\}_{LSF(qd\_anterior)}$ ou $\{a_k\}_{LSF(l-1)}$.

Saída: Os vetores dos coeficientes LPC dos sub-quadros do quadro atual => $\{a_k\}_{sqi}$ ou $\{a_k\}_{(sq\_atual)}$.

Processamento:

- Interpolação linear dos coeficientes LSF por sub-quadro.

- Conversão dos coeficientes LSF do sub-quadro (LSF_{sq} => $\{a_k\}_{LSF(sqi)}$) para os coeficientes LPC do sub-quadro (LPC_{sq}=> $\{a_k\}_{(sqi)}$).

Notação dos parâmetros:

Coeficientes LSF do sub-quadro:

$$\{a_k\}_{LSF(sqi)} => \{a_k\}_{LSF(sq0)}, \{a_k\}_{LSF(sq1)}, \cdots, \{a_k\}_{LSF(sq7)}$$

onde:

$$\{a_k\}_{LSF(sq0)} = (a_{1LSF(sq0)}, a_{2LSF(sq0)}, \cdots, a_{pLSF(sq0)})$$
$$\{a_k\}_{LSF(sq1)} = (a_{1LSF(sq1)}, a_{2LSF(sq1)}, \cdots, a_{pLSF(sq1)})$$
$$\vdots \qquad\qquad \vdots$$
$$\{a_k\}_{LSF(sq7)} = (a_{1LSF(sq7)}, a_{2LSF(sq7)}, \cdots, a_{pLSF(sq7)}).$$

Coeficientes LPC do sub-quadro:

$$\{a_k\}_{(sqi)} => \{a_k\}_{(sq0)}, \{a_k\}_{(sq1)}, \cdots, \{a_k\}_{(sq7)}$$

onde:

$$\{a_k\}_{(sq0)} = (a_{1(sq0)}, a_{2(sq0)}, \cdots, a_{p(sq0)})$$
$$\{a_k\}_{(sq1)} = (a_{1(sq1)}, a_{2(sq1)}, \cdots, a_{p(sq1)})$$
$$\vdots \qquad\qquad \vdots$$
$$\{a_k\}_{(sq7)} = (a_{1(sq7)}, a_{2(sq7)}, \cdots, a_{p(sq7)}).$$

Bloco C110 (Filtro de Análise LP):

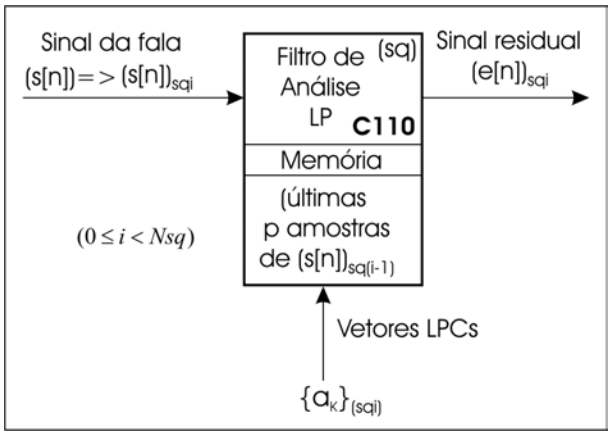

Figura 4.27 – Diagrama de blocos do (processo C110 – Filtro de Análise LP) componente do (processo C100 – bloco de Análise) do codificador WI com indicação dos processamentos e dos parâmetros de entrada e de saída.

Para o processamento do quadro atual são necessárias as amostras em $(s[n]) = (s[n])_{mem} + (s[n]_{qd\_atual}) + (s[n]_{av})$, onde:

$(s[n])_{mem}$ => As últimas p amostras do qd_anterior;

$(s[n]_{qd\_atual}) \Leftrightarrow (s[n]_l)$ => $(s[n])_{sq0}, (s[n])_{sq1}, \cdots, (s[n])_{sq7}$;

$(s[n]_{av})$ => L_a amostras em avanço.

Nesta implementação L_a=120 amostras (6 sub-quadros).

Durante o processamento a memória do filtro sempre armazena as últimas p amostras do sub-quadro anterior, $(s[n])_{sq(i-1)}$.

Entrada:

O bloco C110 recebe o quadro atual $(s[n]_l)$ dividido por sub-quadros $(s[n])_{sqi}$ onde $0 \le i < (Nsq = 8)$ e os vetores LPC $\{a_k\}_{sqi}$.

Na seqüência recebe as amostras em avanço $(s[n]_{av})$ também divididas em sub-quadros.

Na memória:

Na memória do filtro devem estar armazenadas as últimas p amostras do sub-quadro anterior $(s[n])_{sq(i-1)}$. Se o processamento for para o primeiro sub-quadro $(s[n])_{sq1}$, na memória devem estar armazenadas $(s[n])_{mem}$, as últimas p amostras do quadro anterior.

Saída:

O sinal residual $(e[n])_{sqi}$ em correspondência ao processamento (à filtragem do sinal) de cada sub-quadro $(s[n])_{sqi}$.

Processamento:

Para cada sub-quadro de amostras da fala $(s[n])_{sqi}$ do quadro atual $(s[n]_l)$ faz-se:

(1) Conversão de LPC, $\{a_k\}_{sqi}$ com $(1 \le k \le p)$, para os coeficientes do filtro, $\{ec_k\}_{sqi}$ com $(0 \le k \le p)$ onde $\{ec_0 = 1, ec_1 = -a_1, ec_2 = -a_2, \cdots, ec_p = -a_p\}_{sqi}$;

(2) Na Equação (2.9) substitui-se os coeficientes LPC, $\{a_k\}_{sqi}$, pelos coeficientes do filtro, $\{ec_k\}_{sqi}$, que resulta na Equação $e[n] = \sum_{k=0}^{p} ec_k s[n-k]$. Aplica-se esta última equação para cada sub-quadro $(s[n])_{sqi}$ de $(s[n]_{qd\_atual}) \Leftrightarrow (s[n]_l)$ e determina-se cada $(e[n])_{sqi}$;

(3) Repete o item (2) para cada sub-quadro $(s[n])_{sqi}$ de $(s[n]_{av})$. Para a filtragem de cada sub-quadro $(s[n])_{sqi}$ do sinal $(s[n]_{av})$ utiliza-se o último vetor $\{ec_k\}_{sqi}$ do quadro atual, ou seja, $i = 7 \Rightarrow \{ec_k\}_{sq7}$.

(IV) A Estimação e Interpolação do Picth

Bloco C140 (Estimador do Pitch):

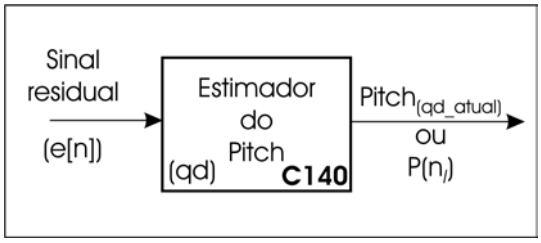

Figura 4.28 – Diagrama de blocos do (processo C140 – Estimador do Pitch) componente do (processo C100 – bloco de Análise) do codificador WI com **i**ndicação dos processamentos e dos parâmetros de entrada e de saída.

Entrada: $(e[n]) = (e[n]_{qd\_atual}) + (e[n]_{av})$ onde:

$(e[n]_{qd\_atual}) \Leftrightarrow (e[n]_l)$ => amostras do quadro atual ($L_q = 160am$);

$(e[n]_{av})$ => amostras do quadro em avanço $L_{av} = 120am$.

Saída: $Pitch_{(qd\_atual)}$ => pitch estimado do quadro atual.

Funções: FACN Equação (4.7) e FCCN Equação (4.8).

Processamento: em 4 janelas mais as regras de decisão. Conforme procedimentos descritos no item 4.3.3.2 Estimação do Pitch e esquematizado nas Figuras 4.13-a e b.

Bloco C150 (Interpolador do Pitch):

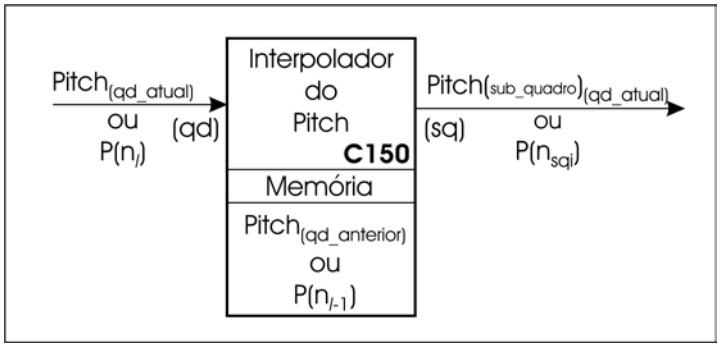

Figura 4.29 – Diagrama de blocos do (processo C150 – Interpolador do Pitch) componente do (processo C100 – bloco de Análise) do codificador WI com indicação dos processamentos e dos parâmetros de entrada e de saída.

Entrada: $Pitch_{(qd\_atual)}$ ou $P(n_l)$ => pitch estimado do quadro atual .

Na memória: Pitch$_{(qd_anterior)}$ ou $P(n_{l-1})$ => pitch estimado do quadro anterior.

Saída: Pitch(sub_quadro)$_{qd\_atual}$ => pitch interpolado a nível de sub-quadro $\{P(n_{sq0}), P(n_{sq1}), \cdots, P(n_{sq7})\}$ para o quadro atual.

Processamento: Interpolação linear => se $|P(n_l) - P(n_{l-1})| < Limiar$

ou Interpolação linear com degrau => se $|P(n_l) - P(n_{l-1})| \geq Limiar$.

Nesta implementação Limiar = 20.

(V) A Localização e Extração das CW's

Bloco C160 (Extração das CW's):

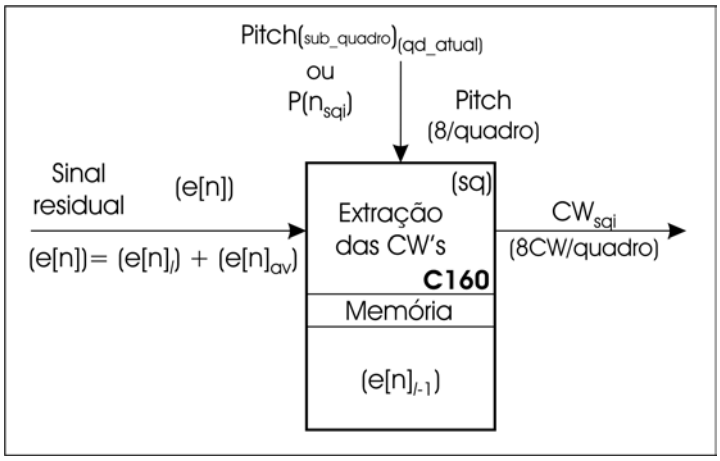

Figura 4.30 – Diagrama de blocos do (processo C160 – Extração das CW's) componente do (processo C100 – bloco de Análise) do codificador WI com indicação dos processamentos e dos parâmetros de entrada e de saída.

Entrada:

$(e[n]) = (e[n]_l) + (e[n]_{av})$ onde:

$(e[n]_l)$ => $(e[n]_{qd\_atual})$ amostras do quadro atual $L_q = 160am$;

$(e[n]_{av})$ => primeiras delta2 amostras do quadro posterior (em avanço).

$P(n_{sqi})$ $(0 \leq i < Nsq)$ => Pitch(sub_quadro) => $\{P(n_{sq0}), P(n_{sq1}), \cdots, P(n_{sq7})\}$.

Na memória:

$(e[n]_{l-1})$ => últimas delta1 amostras do quadro anterior.

Nesta implementação: delta1 = 61 amostras e delta2 = 81 amostras.

Saída:

$CW_{sqi} \Leftrightarrow v(n_{sqi}, \phi) \Leftrightarrow \{A_k(n_{sqi}), B_k(n_{sqi})\}$ $(0 \leq i < Nsq)$. As CW's representadas por seus coeficientes de Fourier.

Nesta implementação Nsq = 8.

Processamento:

(1ª) Localização e extração das CW's

Envolve:

Jee e Jed => janelas de verificação da energia para determinar os limites das CW's;

J_{Ext} => Janela de extração para localização das CW's;

Posição da J_{Ext} em função de $\left(n_{sqi}+\varepsilon, P(n_{sqi})\right)$;

E as CW's representadas no domínio do tempo por:

$$v(n_{sqi}, m) \Rightarrow \textit{Equação}\ (4.14) \quad \textit{ou} \quad v(n_{sqi}, \phi) \Rightarrow \textit{Equação}\ (4.15).$$

(2ª) Cálculo dos coeficientes de Fourier da DTFS => Equação (4.3):

$$CW_{sqi} \Leftrightarrow \left\{A_k(n_{sqi}), B_k(n_{sqi})\right\} \quad (0 \le i < Nsq).$$

CW's representadas no domínio da freqüência por:

$$v(n_{sqi}, m) \Rightarrow \textit{Equação}\ (4.4) \quad \textit{ou} \quad v(n_{sqi}, \phi) \Rightarrow \textit{Equação}\ (4.6).$$

Cada CW fica definida por $\left(n_{sqi}, P(n_{sqi}), \left\{A_k(n_{sqi}), B_k(n_{sqi})\right\}\right)$ onde:

$n_{sqi} \Leftrightarrow$ *posição regular de extração* (considerada como a posição de extração da CW);

$P(n_{sqi}) \Leftrightarrow$ *Pitch* (*comprimento da CW*) $\Rightarrow$ *№ de amostras;*

$\left\{A_k(n_{sqi}), B_k(n_{sqi})\right\} \Leftrightarrow$ *Coeficientes de Fourier da DTFS*, $1 \le k \le \lfloor P(n_{sqi})/2 \rfloor$.

Para mais detalhe veja o item 4.3.3.4 Extração das CW's.

(VI) O Alinhamento das CW's

Bloco C170 (Alinhamento das CW's):

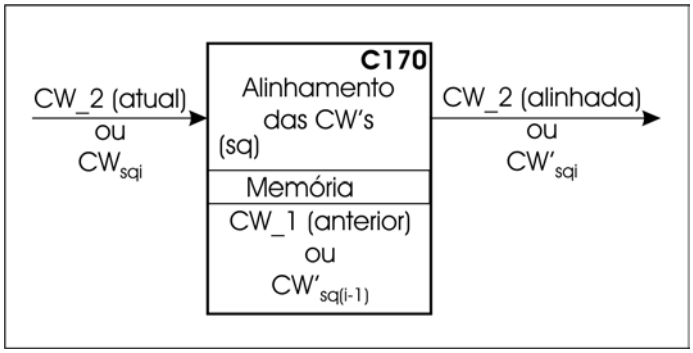

Figura 4.31 – Diagrama de blocos do (processo C170 – Alinhamento das CW's) componente do (processo C100 – bloco de Análise) do codificador WI com indicação dos processamentos e dos parâmetros de entrada e de saída.

Entrada: CW_2 (atual) $(ou\ CW_{sqi}) \Rightarrow$

$$v(n_{sqi}, \phi) \Leftrightarrow \left\{A_k(n_{sqi}), B_k(n_{sqi})\right\} \quad (0 \le i < Nsq).$$

Na memória: CW_1(anterior) $(ou\ CW'_{sq(i-1)}) \Rightarrow$

$$u(n_{sq(i-1)}, \phi) \Leftrightarrow \left\{A'_k(n_{sq(i-1)}), B'_k(n_{sq(i-1)})\right\} (0 \le i < Nsq).$$

Saida: CW_2(atual_alinhada)=> $(ou\ CW'_{sqi}) \Rightarrow$

$$u(n_{sqi},\phi) \Leftrightarrow \left\{A'_k(n_{sqi}), B'_k(n_{sqi})\right\} \quad (0 \le i < Nsq).$$

Subprocessos utilizados:

(1) C171 – Inserção de zeros (para pitch com (sub) múltiplos);

(2) C172 – Truncamento ou acréscimo de zeros para a CW com diferença nos comprimentos menores que um (sub) múltiplo;

(3) C173 – Otimização do critério de alinhamento (Cálculo do deslocamento de fase τ) que alinha a CW_2 com a CW_1. Ver algoritmo no item 4.3.3.5.1 Algoritmo para o Processo C173;

(4) C174 – Modificação dos coeficientes da DTFS - (Alinha a CW_2 (atual) com a CW_1 (anterior));

(5) C175 – Armazena na memória a CW_2 alinhada como CW_1 (anterior) e recebe (ou atualiza) a próxima CW_2 na entrada do processo C170.

(VII) A Normalização da Potência das CW's

Bloco C180 (Cálculo da potência das CW's):

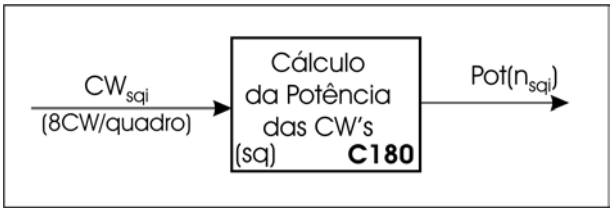

Figura 4.32 – Diagrama de blocos do (processo C180 – Cálculo da Potência das CW's) componente do (processo C100 – bloco de Análise) do codificador WI com indicação dos processamentos e dos parâmetros de entrada e de saída.

Entrada: CW (atual) => $CW_{sqi} \Leftrightarrow v(n_{sqi},\phi) \Leftrightarrow \left\{A_k(n_{sqi}), B_k(n_{sqi})\right\} \quad (0 \le i < Nsq)$.

Saída: $Pot(n_{sqi})$

Processamento: Aplica-se a Equação (4.30).

Bloco C190 (Normalização da potência das CW's):

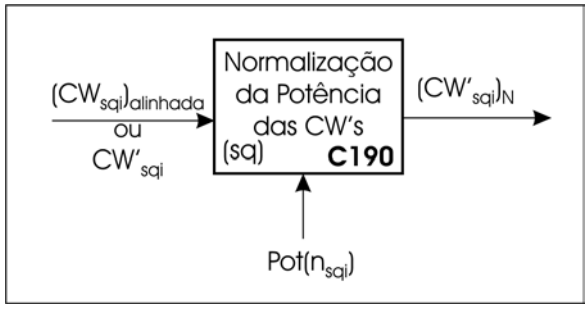

Figura 4.33 – Diagrama de blocos do (processo C190 – Normalização da Potência das CW's) componente do (processo C100 – bloco de Análise) do codificador WI com indicação dos processamentos e dos parâmetros de entrada e de saída.

Entrada:

CW (atual) =>

$$(CW_{sqi})_{alinhada}\ (ou\ CW'_{sqi}) \Leftrightarrow u(n_{sqi},\phi) \Leftrightarrow \left\{A'_k(n_{sqi}), B'_k(n_{sqi})\right\}(0 \le i < Nsq)$$

E a potência da CW => $Pot(n_{sqi})$.

Saída:

CW atual normalizada em potência - CW $(atual)_N$ =>

$$(CW_{sqi})_N \Leftrightarrow u_N(n_{sqi},\phi) \Leftrightarrow \left\{A_{Nk}(n_{sqi}), B_{Nk}(n_{sqi})\right\}(0 \le i < Nsq)\,.$$

Processamento: Aplica-se a Equação (4.32).

4.3.4 O Estágio da Síntese

No esquema do codificador WI, mostrado na Figura 4.4, a função do bloco de síntese (processo D200), é converter os parâmetros recebidos do codificador (processo C100 – bloco de análise) em sinal da fala (sinal reconstruído). Os valores dos parâmetros que chegam ao decodificador, em geral, não são exatamente os mesmos que foram enviados pelo estágio de análise C100 devido a compressão e a transmissão. Apesar da simulação neste trabalho envolver somente a transmissão dos parâmetros sem compressão e, sem considerar a influência do canal de transmissão, será adotada no estágio de síntese, a notação "^" para os parâmetros originais descritos no estágio de análise, indicando que os parâmetros foram codificados, transmitidos e decodificados. Desta forma, os parâmetros ficam com uma notação pronta para os trabalhos futuros que envolvam a camada interna, ou camada de compressão, e o canal de transmissão, considerando a influência de ruídos. Assim, os parâmetros recebidos pelo decodificador em correspondência com os parâmetros transmitidos, são os coeficientes LSF's, o pitch, a potência e as formas de ondas características (CW's normalizadas) que estão relacionados no item 4.3.3.7 deste capítulo e são descritos por:

(1) Os coeficientes LSF's $\{\hat{a}_k\}_{LSF(qdl)}$ ou $\{\hat{a}_k\}_{LSF(qd\_Atual)}$ =>> (1 vetor(Np ordenadas)/quadro);

(2) O pitch $\hat{P}(n_l)$, $\hat{P}(qd\_Atual)$ ou $\hat{P}_{qdl}$ ==========>> (1 escalar/quadro);

(3) A potência das CW's $(\hat{CW}_{sqi})_N$ => ($\hat{Pot}(n_{sqi})$) ou $\hat{Pot}(sq\_Atual)$ =>> (8escalares/quadro);

(4) As CW's $\hat{u}_N(n_{sqi},\phi) \Leftrightarrow \left\{\hat{A}_{Nk}(n_{sqi}), \hat{B}_{Nk}(n_{sqi})\right\}$ ==========>> (8 conjuntos/quadro).

Na implementação do decodificador estes parâmetros são lidos a partir do arquivo "par_wi.c" que foi gerado durante o processo de análise conforme o item *4.3.3.7* deste capítulo, *Parâmetros de Saída do Processo C100 (Bloco de Análise).*

O estágio de síntese consiste em vários processos distintos que são esquematizados na Figura 4.34 onde o bloco de síntese (processo D200) está expandido. O diagrama esquemático representa uma implementação típica para o decodificador WI convencional. A descrição de cada um dos processos envolvidos é introduzida nas próximas subseções.

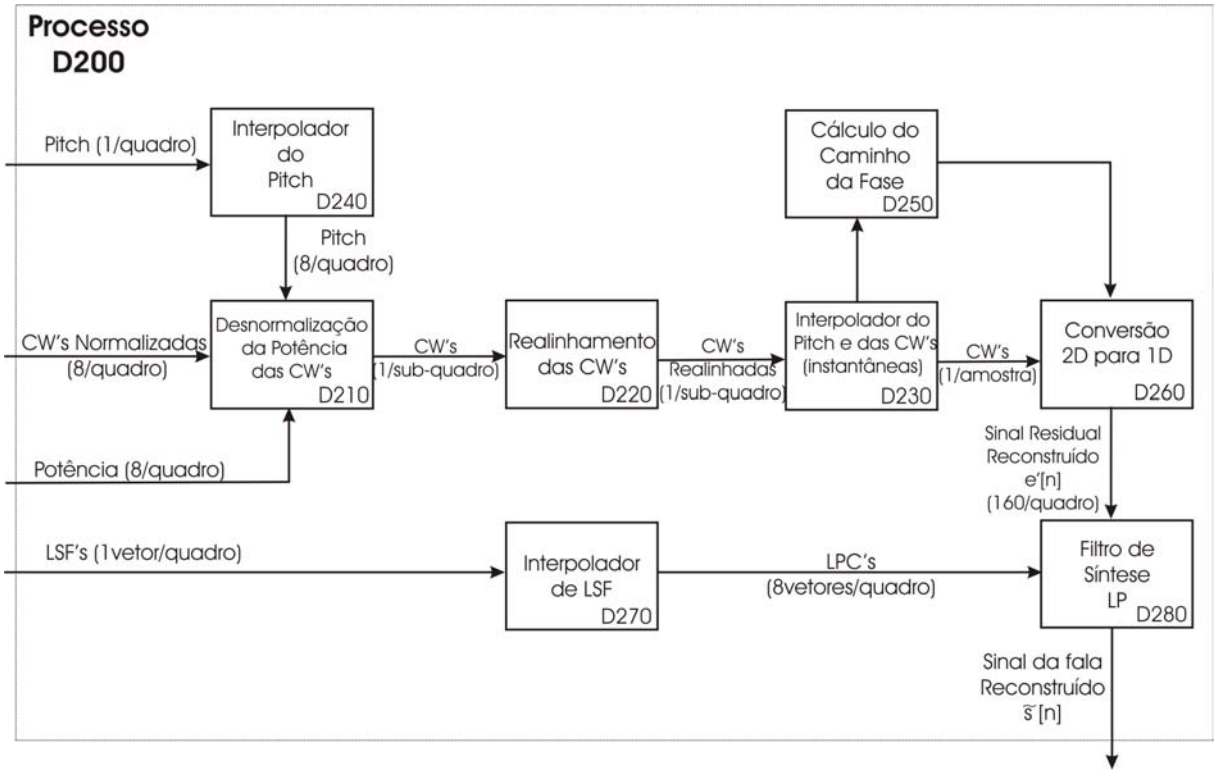

Figura 4.34 – Diagrama de blocos de síntese expandido (processo D200) do codificador WI [14]. Na Figura 4.45 o bloco de síntese expandido (processo D200) é apresentado com a inclusão dos parâmetros de entrada e de saída de cada bloco.

4.3.4.1 Interpolador do Pitch e Interpolador de LSF

O processo D240 – *Interpolador do pitch* é idêntico ao processo C150 – *Interpolador do pitch* (no codificador) descrito no item 4.3.3.3, Interpolação do Pitch. Assim, o processo D240 recebe os valores do pitch por quadro, $pitch_{(quadro\ atual)} \Leftrightarrow \hat{P}(n_l)$ e $pitch_{(quadro\ anterior)} \Leftrightarrow \hat{P}(n_{l-1})$, e faz a interpolação por sub-quadro resultando nos valores de pitch por sub-quadro $\{\hat{P}(n_{sq0}), \hat{P}(n_{sq1}), \ldots, \hat{P}(n_{sqi}), \ldots, \hat{P}(n_{sq7})\}$ que tornam-se disponíveis para o processo D210 – Desnormalização da Potência das CW's. Mais detalhes foram descritos no item 4.3.3.3, deste capítulo.

O processo D270 – *Interpolador de LSF* também é idêntico ao processo C120 – *Interpolador de LSF* (no codificador) descrito no item 4.3.3.1 deste capítulo, Análise com Predição Linear. Assim o processo D270 recebe os valores dos coeficientes LSF por quadro, $\{\hat{a}_k\}_{LSF(l)}$ $(quadro\ atual)$ e $\{\hat{a}_k\}_{LSF(l-1)}$ $(quadro\ anterior)$, faz a interpolação por sub-quadro, resultando em LSF(sub-quadros) e em seguida, os coeficientes LSF (sub-quadros) são convertidos para LPC (sub- quadro), que tornam-se disponíveis para o processo D280 – Filtro de síntese LP. Mais detalhes foram descritos no item 4.3.3.1, deste capítulo.

4.3.4.2 Desnormalização da Potência das CW's

Para desnormalizar as CW's (processamento por sub-quadro), o processo D210 – *Desnormalização da Potência das CW's* recebe o pitch por sub-quadro $\hat{P}(n_{sqi})$, a potência $\hat{P}ot(n_{sqi})$ e a CW normalizada $(\hat{C}W_{sqi})_N$. O valor do pitch que define o comprimento da forma de onda característica em número de amostras, permite também a determinação do comprimento da $(\hat{C}W_{sqi})_N$ que chegou ao decodificador em número de coeficientes de Fourier, que é obtido por $\lfloor \hat{P}(n_{sqi})/2 \rfloor$ coeficientes da DTFS (ou harmônicas), onde $\lfloor \hat{P}(n_{sqi})/2 \rfloor = \hat{P}(n_{sqi})/2,\ se\ \hat{P}(n_{sqi})$ for par, e $\lfloor \hat{P}(n_{sqi})/2 \rfloor = (\hat{P}(n_{sqi})-1)/2,\ se\ \hat{P}(n_{sqi})$ for ímpar. À $(\hat{C}W_{sqi})_N \Leftrightarrow \hat{u}_N(n_{sqi}, \phi)$ representada pelos seus coeficientes da DTFS, $\{\hat{A}_{Nk}(n_{sqi}), \hat{B}_{Nk}(n_{sqi})\}$ $(0 \le i < Nsq)$, aplica-se inverso do que foi realizado pela Equação (4.32), obtendo-se a $\hat{C}W_{sqi} \Leftrightarrow \hat{u}(n_{sqi}, \phi)$ desnormalizada e representada pelos seus coeficientes da DTFS, $\{\hat{A}_k(n_{sqi}), \hat{B}_k(n_{sqi})\}$ $(0 \le i < Nsq)$, onde:

$$\begin{aligned} \hat{A}_k(n_{sqi}) &= \hat{A}_{Nk}(n_{sqi})\sqrt{\hat{P}ot(n_{sqi})}, \\ \hat{B}_k(n_{sqi}) &=, \hat{B}_{Nk}(n_{sqi})\sqrt{\hat{P}ot(n_{sqi})}. \end{aligned} \tag{4.33}$$

A Figura 4.35 representa o processo D210.

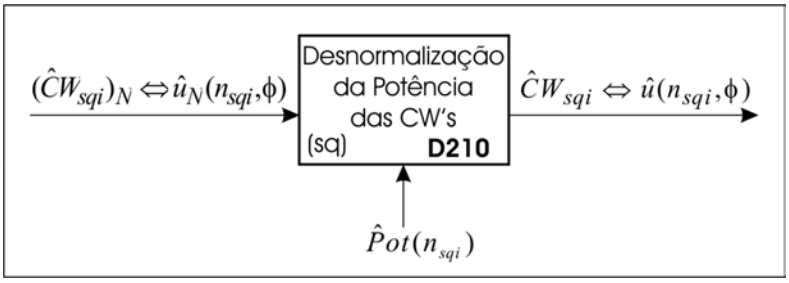

Figura 4.35 – Diagrama de blocos do (processo D210 – Desnormalização da Potência das CW's) componente do (processo D200 – bloco de Síntese -Decodificador) no codificador WI com **i**ndicação dos parâmetros de entrada e de saída.

4.3.4.3 Realinhamento das CW's

A codificação e a transmissão das CW's podem mudar a forma da representação bidimensional. Assim, um procedimento de realinhamento é usualmente aplicado nas formas de ondas recebidas no decodificador para assegurar que elas continuem alinhadas. O processo D220 – Realinhamento das CW's é idêntico ao processo C170 – Alinhamento das CW's (no codificador) descrito no item 4.3.3.5 deste capítulo, Alinhamento das CW's. Assim o processo D220 recebe a $\hat{C}W_{sqi} \Leftrightarrow \hat{u}(n_{sqi},\phi)$ desnormalizada, representada pelos seus coeficientes da DTFS, $\left\{\hat{A}_k(n_{sqi}),\hat{B}_k(n_{sqi})\right\}(0 \le i < Nsq)$, e obtém-se a $\hat{C}W'_{sqi} \Leftrightarrow \hat{u}'(n_{sqi},\phi)$ realinhada, representada pelos seus coeficientes da DTFS, $\left\{\hat{A}'_k(n_{sqi}),\hat{B}'_k(n_{sqi})\right\}$ $(0 \le i < Nsq)$ que fica disponível para o processo D230. Mais detalhes são descritos no item 4.3.3.5 deste capítulo, Alinhamento das CW's. A Figura 4.36 representa o processo D220.

Figura 4.36 – Diagrama de blocos do (processo D220 – Realinhamento das CW's) componente do (processo D200 – bloco de Síntese -Decodificador) do codificador WI com **i**ndicação dos parâmetros de entrada e de saída.

4.3.4.4 Geração das CW's e Pitch instantâneos

Nesta fase as CW's, $\hat{CW}'_{sqi} \Leftrightarrow \hat{u}'(n_{sqi},\phi)$ com $0 \le i < Nsq$, estão prontas para gerar a superfície bidimensional que possibilita a reconstrução do sinal residual para o quadro atual. Para isto a evolução das formas de ondas entre dois instantes de amostragem sucessivos no tempo, ao longo do eixo n, pode ser aproximada por interpolação utilizando as $(\hat{CW's})_{am}$ entre duas CW's extraídas, $\hat{CW}'_{sq(i-1)}$ (anterior) e $\hat{CW}'_{sqi}$ (atual) de sub-quadros adjacentes. Assim, para cada instante de amostragem n_a ao longo do eixo de evolução das formas de ondas existirá uma CW interpolada, $(\hat{CW}')_{am} \Leftrightarrow \hat{u}'(n_a,\phi)$ *onde* $0 \le a \le Lsq$ e $n_{sq(i-1)} \le n_a \le n_{sqi}$. Isto resulta em uma representação bidimensional das CW's praticamente contínua ao longo do eixo dos tempo n. A representação utilizando-se as séries de Fourier efetivamente faz as formas de ondas contínuas ao longo do eixo abstrato das fases, desde que a Equação (4.6) possa permitir o cálculo de $\hat{u}'(n_a,\phi)$ para qualquer fase $(0 \le \phi < 2\pi)$. Portanto, as formas de ondas extraídas, $(\hat{CW's})_{sqi}$, em conjunto com as formas de ondas interpoladas, $(\hat{CW's})_{am}$, podem ser imaginadas como uma superfície de formas de ondas características contínuas, $\hat{u}'(n,\phi)$. O sinal residual pode, portanto, ser reconstruído usando esta superfície. Em cada instante de amostragem n_a ao longo do eixo n é necessário determinar o valor do pitch, $\hat{P}(n_a)$, que corresponde ao comprimento de cada CW, $(\hat{CW}')_{am}$, a ser interpolada no respectivo instante, sendo esse valor, denominado de pitch instantâneo e a $(\hat{CW}')_{am} \Leftrightarrow \hat{u}'(n_a,\phi)$ correspondente, denominada de CW instantânea.

A interpolação do pitch, $\hat{P}(n_a)$, e das CW's, $(\hat{CW}')_{am} \Leftrightarrow \hat{u}'(n_a,\phi)$, instantâneos (ou por amostra) é realizada pelo processo D230 –*Interpolador do pitch e das CW's (instantâneos)* mostrado na Figura 4.37.

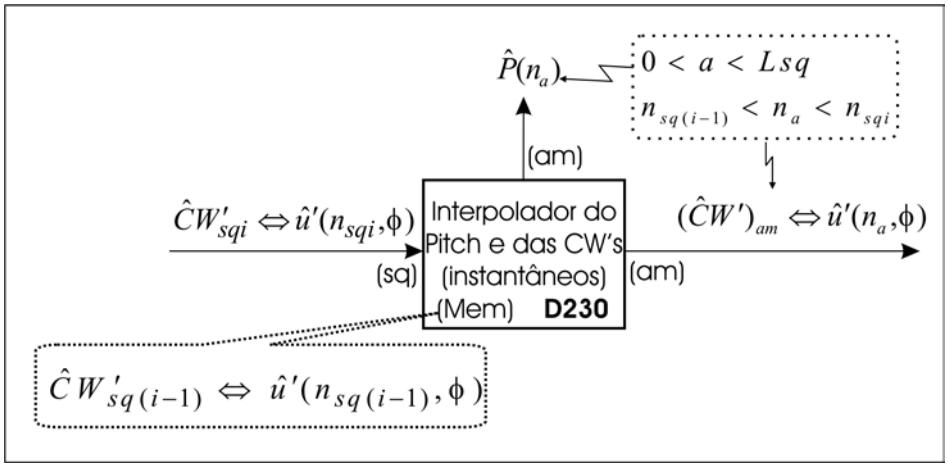

Figura 4.37 – Diagrama de blocos do (processo D230 – Interpolador do pitch e das CW's) componente do (processo D200 – bloco de Síntese - Decodificador) no codificador WI com indicação dos parâmetros de entrada e de saída. Na entrada recebe as $\hat{C}W'_{sqi}$ por sub-quadro que resultam na saída em $(\hat{C}W')_{am}$ e $\hat{P}(n_a)$, respectivamente, as formas de ondas e o pitch interpolados por amostra. Na memória (Mem) fica armazenada a $\hat{C}W'_{sq(i-1)}$ do sub-quadro anterior.

Para realizar a interpolação, as CW's, $\hat{C}W'_{sq(i-1)}$ (anterior) e $\hat{C}W'_{sqi}$ (atual) devem ter a mesma dimensão. Quando as CW's têm dimensões diferentes é necessário aplicar um pré-processamento sobre elas no domínio dos coeficientes de Fourier (ou da freqüência) tais como inserção de zeros e/ou acréscimo (ou apêndice) de zeros, de forma que elas tenham a mesma dimensão (ou a mesma quantidade de harmônicas), antes da interpolação instantânea das CW's. Estes procedimentos permitem a aplicação uniforme de um mesmo critério para a interpolação, o que facilita a implementação. Para facilitar a compreensão e a implementação, o processo D230 *(Interpolador do pitch e das CW's*), esquematizado na Figura 4.39, é dividido nos processos D231, D232, D233, D234, D235 e D236 em [14], que são descritos a seguir:

Processo D231 (*Inserção de zeros*) – Este processo é utilizado quando a diferença entre os comprimentos das CW's, CW atual e CW anterior, é maior ou igual a um múltiplo do comprimento da CW mais curta. O processo é utilizado se o pitch tiver múltiplos ou sub-múltiplos. A CW menor é estendida periodicamente em um número inteiro de vezes no processo D231 para aproximá-la ao máximo do comprimento da CW maior. Mais detalhes sobre o processo de *Inserção de zeros* são descritos no item 4.3.3.5 deste capítulo, Alinhamento das CW's. Após os zeros serem inseridos, as CW's são enviadas ao processo D232. Os zeros inseridos pelo processo D231 são deixados nas CW's que passarão pelo processo D260. Assim, o comprimento da CW's e conseqüentemente o pitch após a inserção de zeros ficam modificados. Por uma questão de consistência o valor do pitch é atualizado e

então enviado para o processo de interpolação do pitch, processo D235 *(Interpolador linear do pitch)*, através das variáveis $\hat{P}_{rf}(n_{sqi})$ e $\hat{P}_{rf}(n_{sq(i-1)})$ que representam os valores de pitch de referência para a interpolação do pitch instantâneo. Desta forma, o pitch instantâneo que vai para o processo D250 *(Estimação do caminho da fase)* coincidirá com o comprimento da CW instantânea que vai para o processo D260 *(Conversão 2D para 1D)*.

Processo D232 (*Acréscimo de zeros para a CW mais curta*) – Este é o último processamento efetuado sobre as CW's, CW atual ou CW anterior, para que elas tenham o mesmo comprimento antes da aplicação do processo D233- *(Interpolação das CW's)*. Se as CW's tiverem comprimentos diferentes, mas a diferença é menor do que um múltiplo ou sub-múltiplo do pitch, aplica-se o acréscimo (ou apêndice) de zeros no domínio espectral (aos coeficientes DTFS da CW mais curta). Mais detalhes sobre o *Acréscimo (preenchimento ou apêndice) espectral com zeros* são descritos no item 4.3.3.5 deste capítulo, Alinhamento das CW's.

Processo D233 (*Interpola* $(\hat{CW}_I's)_{am}$ *entre a* $\hat{CW}'_{sq(i-1)}$ *e a* $\hat{CW}'_{sqi}$) – A Figura 4.38 mostra o diagrama esquemático para a interpolação das $(\hat{CW}_I's)_{am}$. Este processo faz a interpolação das $(\hat{CW}_I's)_{am}$ no sub-quadro atual entre duas CW's de sub-quadros consecutivos que têm o mesmo comprimento. O processo recebe as CW's de dois sub-quadros adjacentes, a CW do sub-quadro anterior, $\hat{CW}'_{sq(i-1)}$, e a CW do sub-quadro atual, $\hat{CW}'_{sqi}$, e aplica-se a interpolação linear aos coeficientes de Fourier, que é equivalente a interpolação linear no domínio do tempo entre duas CW's [14]. Assim para as CW's representadas por $\hat{CW}'_{sqi} \Leftrightarrow \hat{u}'(n_{sqi}, m) \Leftrightarrow \{\hat{A}'_k(n_{sqi}), \hat{B}'_k(n_{sqi})\}$ e $\hat{CW}'_{sq(i-1)} \Leftrightarrow \hat{u}'(n_{sq(i-1)}, m) \Leftrightarrow \{\hat{A}'_k(n_{sq(i-1)}), \hat{B}'_k(n_{sq(i-1)})\}$ que correspondem respectivamente a $\mathrm{CW\_}2B$ e $\mathrm{CW\_}1B$ na Figura 4.39, a Equação da interpolação da $(\hat{CW}_I')_{am}$ na posição n_a representada por $(\hat{CW}_I')_{am} \Leftrightarrow \hat{u}_I'(n_a, m) \Leftrightarrow \{\hat{A}'_{Ik}(n_a), \hat{B}'_{Ik}(n_a)\}$ com $(0 < a < Lsq)$ e $(n_{sq(i-1)} < n_a < n_{sq(i-1)})$, torna-se,

$$\hat{A}'_{Ik}(n_a) = \frac{n_{sqi} - n_a}{n_{sqi} - n_{sq(i-1)}} \hat{A}'_k(n_{sq(i-1)}) + \frac{n_a - n_{sq(i-1)}}{n_{sqi} - n_{sq(i-1)}} \hat{A}'_k(n_{sqi})$$

$$\hat{B}'_{Ik}(n_a) = \frac{n_{sqi} - n_a}{n_{sqi} - n_{sq(i-1)}} \hat{B}'_k(n_{sq(i-1)}) + \frac{n_a - n_{sq(i-1)}}{n_{sqi} - n_{sq(i-1)}} \hat{B}'_k(n_{sqi}) \qquad (4.34)$$

onde $(1 \le k \le \lfloor \hat{P}(n_{sqi})/2 \rfloor$.

Após o cálculo, todas as CW's interpoladas instantâneas para o sub-quadro atual, $(\hat{CW}_I's)_{am}$, ficam disponíveis para o processo D234 *(Truncamento e ajuste da potência (se necessário))*.

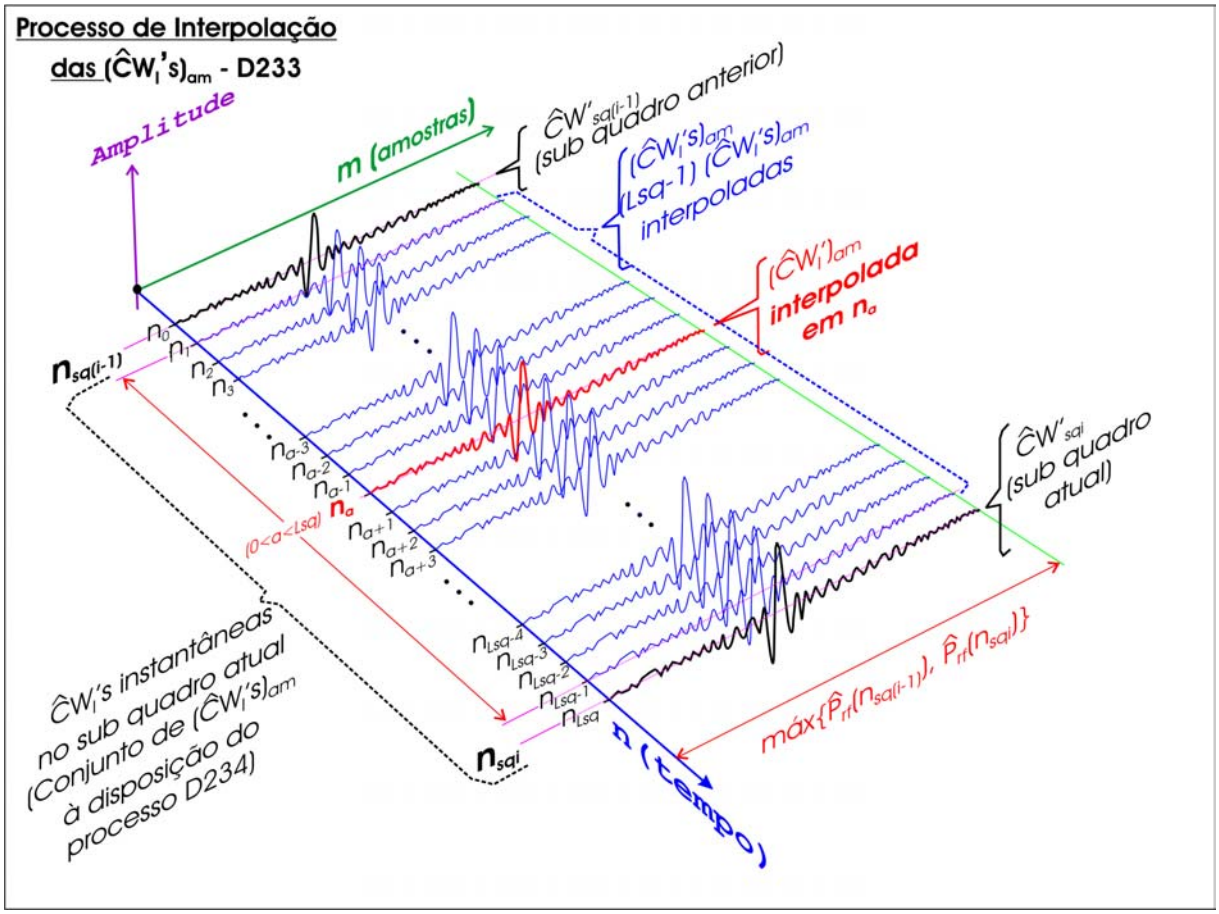

Figura 4.38 – Diagrama esquemático para o processo da interpolação instantânea das CW's - Processo D233 (*Interpola* $(\hat{CW}'s)_{am}$ *entre a* $\hat{CW}'_{sq(i-1)}$ *e a* $\hat{CW}'_{sqi}$).

Processo D234 *(Truncamento e ajuste da potência (se necessário))* – Este processo recebe as CW's interpoladas instantâneas, $(\hat{CW}_I's)_{am}$, a partir do processo D233, todas com o mesmo comprimento, igual ao maior valor entre $\hat{P}_{rf}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi})$. Recebe também os valores do pitch instantâneo $\hat{P}(n_a)$ (com valores inteiros), a partir do processo D235. Para cada valor de n_a com $(0 < a < Lsq)$ $(n_{sq(i-1)} < n_a < n_{sq(i-1)})$ compara os valores de $\hat{P}(n_a)$ com o comprimento da correspondente $(\hat{CW}_I')_{am} \Leftrightarrow \hat{u}_I'(n_a, m) \Leftrightarrow \{\hat{A}'_{Ik}(n_a), \hat{B}'_{Ik}(n_a)\}$. Se o valor de $\hat{P}(n_a)$ for menor do que o comprimento da $\hat{CW}'_{am} \Leftrightarrow \hat{u}'(n_a, m)$ trunca-se os coeficientes (truncamento espectral) correspondentes às freqüências mais altas de $\{\hat{A}'_{Ik}(n_a), \hat{B}'_{Ik}(n_a)\}$ que excederem ao valor $\lfloor \hat{P}(n_a)/2 \rfloor$, reduzindo o comprimento da correspondente $(\hat{CW}_I')_{am} \Leftrightarrow \hat{u}_I'(n_a, m)$ para o valor de $\hat{P}(n_a)$. O outro caso que pode ocorrer é $\hat{P}(n_a)$ ter o mesmo valor que o comprimento da $(\hat{CW}_I')_{am} \Leftrightarrow \hat{u}_I'(n_a, m)$. Neste

caso a $(\hat{CW}'_I)_{am} \Leftrightarrow \hat{u}'_I(n_a, m)$ permanece inalterada. No caso quando ocorre truncamento de coeficientes, a CW instantânea perde potência. Para evitar a perda de potência neste processo, que pode causar distorções no nível do sinal reconstruído, calcula-se a potência correspondente aos coeficientes truncados (ou harmônicas desprezadas) e faz-se uma compensação (ou distribuição) desta potência sobre os coeficientes remanescentes na CW instantânea após o truncamento. Após realizar as operações de truncamento e ajuste da potência com todas CW's interpoladas instantâneas do sub-quadro atual, as CW's resultantes, denominadas de CW's instantâneas (ou CW por amostra) tornam-se disponíveis para a conversão do sinal bidimensional, *superfície das ondas características* $\hat{u}(n,\phi)$, para o sinal unidimensional, *sinal residual*, no processo D260 *(Conversão 2D para 1D)*. Mais detalhes sobre o truncamento espectral são descritos no item 4.3.3.5 - Alinhamento das CW's.

Processo D235 *(Interpolador linear do pitch)* – Neste processo, $\hat{P}_{rf}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi})$, denominados de pitch de referência dos sub-quadros anterior e atual, recebem $\hat{P}(n_{sq(i-1)})$ e $\hat{P}(n_{sqi})$, os valores de pitch das CW's, $\hat{CW}'_{sq(i-1)}$ e $\hat{CW}'_{sqi}$ respectivamente, diretamente a partir da entrada do processo D230 ou do valor corrigido do pitch quando ocorrer a inserção de zeros no processo D231, como indicado na Figura 4.39. A partir dos valores $\hat{P}_{rf}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi})$, o processo D235 realiza a interpolação linear do pitch instantâneo $\hat{P}_f(n_a)$ para o sub-quadro atual, utilizando-se a Equação (4.9) reescrita aqui como:

$$\hat{P}_f(n_a) = \frac{[(n_{sqi} - n_a)\hat{P}_{rf}(n_{sq(i-1)}) + (n_a - n_{sq(i-1)})\hat{P}_{rf}(n_{sqi})]}{(n_{sqi} - n_{sq(i-1)})} \qquad (4.35)$$

$$\text{com} \quad (0 < a < Lsq) \quad \text{e} \quad (n_{sq(i-1)} < n_a < n_{sq(i-1)}).$$

Os valores de $\hat{P}_f(n_a)$ são calculados com ponto flutuante e enviados para o processo D250 para tornar mais preciso o cálculo do caminho da fase, e são arredondados para inteiros, tornando-se $\hat{P}(n_a)$, e enviados para o processo D234 para ajustar o comprimentos das CW's que sofreram acréscimo de zeros no processo D232.

Processo D236 *(Atraso por um sub-quadro)* – Após a interpolação das $(\hat{CW's})_{am}$ do sub-quadro atual, o processo D236 atualiza as variáveis, CW_1 e CW_2 (Figura 4.39), no início do processo D230. A CW_1 na memória recebe os coeficientes da DTFS da CW_{sqi} (CW_atual) que passa a ser CW_anterior, e a CW_2 na entrada do processo D230 recebe a próxima CW, $CW_{sq(i+1)}$, que passa a ser a CW_atual. Neste instante, o processo D230 fica

pronto para a interpolação das $(\hat{CW}'s)_{am}$ do próximo sub-quadro. As CW_1 e CW_2 ficam disponíveis para o processo D231 na seqüência.

Procedimentos para a implementação do processo D230: Na implementação, o processo D230 (*Interpolador do pitch e das CW's (instantâneos)*) é executado em um dos três módulos (módulo 1, módulo 2 ou módulo 3). Na escolha do módulo é utilizado o parâmetro C, definido na Equação (4.24) e na Equação (4.25) e reescritas aqui substituindo $P(n_{sqi})$ e $P(n_{sq(i-1)})$ por $\hat{P}(n_{sqi})$ e $\hat{P}(n_{sq(i-1)})$, respectivamente, onde:

$$C = \frac{\hat{P}(n_{sqi})}{\hat{P}(n_{sq(i-1)})} \tag{4.36}$$

para $\hat{P}(n_{sqi}) \geq \hat{P}(n_{sq(i-1)})$ (múltiplos) ou na Equação (4.37)

$$C = \frac{\hat{P}(n_{sq(i-1)})}{\hat{P}(n_{sqi})} \tag{4.37}$$

para $\hat{P}(n_{sqi}) \leq \hat{P}(n_{sq(i-1)})$ (submúltiplos). Assim os módulos são definidos como:

a) **<u>Módulo 1</u>**: A $\hat{CW}'_{sqi}$ e a $\hat{CW}'_{sq(i-1)}$ têm a mesma dimensão (C=1). Este módulo envolve os processos D233, D235 e D236, onde:
   - Processo D233 -> Interpola as Lsq CW's instantâneas entre a CW_sq_anterior e a CW_sq_atual;
   - Processo D235 -> Faz a interpolação linear do pitch;
   - Processo D236 -> Atualiza as CW's, a memória recebe os coeficientes da DTFS da CW atual como CW_anterior, e a CW_atual recebe os coeficientes da próxima CW.

   <u>Procedimentos no Módulo 1</u>:

   (1) $CW\_1 \Leftarrow \hat{CW}'_{sq(i-1)}$ com $\hat{P}(n_{sq(i-1)})$ e $CW\_2 \Leftarrow \hat{CW}'_{sqi}$ com $\hat{P}(n_{sqi})$.
   (2) $\hat{P}_{rf}(n_{sq(i-1)}) \Leftarrow \hat{P}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi}) \Leftarrow \hat{P}(n_{sqi})$.
   (3) $CW\_1B \Leftarrow CW\_1$ e $CW\_2B \Leftarrow CW\_2$.
   (4) Executa o processo D233
      - Recebe CW_1B e CW_2B;
      - Calcula as $(\hat{CW}'_I s)_{am}$ (interpolação – CW's com a mesma dimensão).
   (5) As $(\hat{CW}'s)_{am}$ (mesma dimensão) <= (recebem) as $(\hat{CW}'_I s)_{am}$ (mesma dimensão);

As $(\hat{C}W's)_{am}$ tornam-se disponíveis para D260.

(6) Executa o processo D235:
- Recebe $\hat{P}_{rf}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi})$;
- Calcula os valores dos $\hat{P}_f(n_a)$;
- Os valores dos $\hat{P}_f(n_a)$ tornam-se disponíveis para D250.

(7) Executa o processo D236:

Atualiza a CW_1 e a CW_2:
- CW_1 <= $\hat{C}W'_{sqi}$ (CW_atual);
- CW_2 <= $\hat{C}W'_{sq(i+1)}$ (CW_posterior);

A CW_1 e a CW_2 tornam-se disponíveis para o próximo processo, D231:
- CW_1 como (CW_anterior);
- CW_2 como (CW_atual).

(8) Fim do Módulo 1.

b) **Módulo 2**: A $\hat{C}W'_{sqi}$ e a $\hat{C}W'_{sq(i-1)}$ têm dimensões diferentes ($1 < C < 2$) (sem múltiplos ou submúltiplo do pitch). Envolve os processos D232, D233, D234, D235 e D236:
- Processo D232 -> Faz o acréscimo de zeros no domínio espectral (coeficientes da DTFS);
- Processo D233 -> Interpola as Lsq CW's instantâneas entre a CW_sq_anterior e a CW_sq_atual;
- Processo D234 -> Truncamento de zeros no domínio espectral (coeficientes da DTFS) e ajuste da potência;
- Processo D235 -> Interpolação linear do pitch;
- Processo D236 -> Atualiza as CW's, a memória recebe os coeficientes da DTFS da CW atual como CW_anterior, e a CW_atual recebe os coeficientes da próxima CW.

Procedimentos no Módulo 2:

(1) $CW\_1 \Leftarrow \hat{C}W'_{sq(i-1)}$ com $\hat{P}(n_{sq(i-1)})$ e $CW\_2 \Leftarrow \hat{C}W'_{sqi}$ com $\hat{P}(n_{sqi})$.

(2) $\hat{P}_{rf}(n_{sq(i-1)}) \Leftarrow \hat{P}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi}) \Leftarrow \hat{P}(n_{sqi})$.

(3) $CW\_1A \Leftarrow CW\_1$ e $CW\_2A \Leftarrow CW\_2$.

(4) Executa o processo D232
- Recebe CW_1A e CW_2A;
- Faz o acréscimo de zeros para a CW mais curta => CW_1B e CW_2B;
- CW_1B e CW_2B tornam-se disponíveis para o próximo processo, D233.

(5) Executa o processo D233
- Recebe CW_1B e CW_2B;
- Calcula as $(\hat{C}W'_I s)_{am}$ (interpolação – CW's com a mesma dimensão);

- As $(\hat{C}W_1's)_{am}$ tornam-se disponíveis para D234.

(6) Executa o processo D235:
- Recebe $\hat{P}_{rf}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi})$;
- Calcula os valores de $\hat{P}_f(n_a)$ e $\hat{P}(n_a)$;
- Os valores de $\hat{P}_f(n_a)$ tornam-se disponíveis para D250 e os de $\hat{P}(n_a)$ para D234.

(7) Executa o processo D234:
- Recebe $\hat{P}(n_a)$ e as $(\hat{C}W_1's)_{am}$;
- Calcula as $(\hat{C}W's)_{am}$;
- As $(\hat{C}W's)_{am}$ tornam-se disponíveis para D260.

(8) Executa o processo D236:

Atualiza a CW_1 e a CW_2:
- CW_1 <= $\hat{C}W'_{sqi}$ (CW_atual);
- CW_2 <= $\hat{C}W'_{sq(i+1)}$ (CW_posterior);

A CW_1 e a CW_2 tornam-se disponíveis para o próximo processo, D231:
- CW_1 como (CW_anterior);
- CW_2 como (CW_atual).

(9) Fim do <u>Módulo 2</u>.

c) **<u>Módulo 3</u>**: A $\hat{C}W'_{sqi}$ e a $\hat{C}W'_{sq(i-1)}$ possuem dimensões diferentes e ocorre múltiplos ou submúltiplo do pitch ($C \geq 2$). Envolve os processos D231, D232, D233, D234, D235 e D236:
- Processo D231 -> Inserção de zeros no domínio espectral (coeficientes da DTFS);
- Processo D232 -> Acréscimo de zeros no domínio espectral (coeficientes da DTFS);
- Processo D233 -> Interpola as Lsq CW's instantâneas entre a CW_sq_anterior e a CW_sq_atual;
- Processo D234 -> Truncamento de zeros no domínio espectral (coeficientes da DTFS) e ajuste da potência;
- Processo D235 -> Interpolação linear do pitch;
- Processo D236 -> Atualiza as CW's, a memória recebe os coeficientes da DTFS da CW atual como CW_anterior, e a CW_atual recebe os coeficientes da próxima CW.

<u>Procedimentos no Módulo 3</u>:

(1) $CW\_1 \Leftarrow \hat{C}W'_{sq(i-1)}$ com $\hat{P}(n_{sq(i-1)})$ e $CW\_2 \Leftarrow \hat{C}W'_{sqi}$ com $\hat{P}(n_{sqi})$.

(2) Executa o processo D231:
- Recebe CW_1 e CW_2;
- Faz a inserção de zeros para a CW mais curta => resulta em:

CW_1A (com $\hat{P}_{rf}(n_{sq(i-1)})$)

e CW_2A (com $\hat{P}_{rf}(n_{sqi})$);

- CW_1A e CW_2A tornam-se disponíveis para o próximo processo, D232;
- $\hat{P}_{rf}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi})$ tornam-se disponíveis para D235.

(3) Executa o processo D232:
- Recebe CW_1A e CW_2A;
- Faz o acréscimo de zeros para a CW mais curta => CW_1B e CW_2B;
- CW_1B e CW_2B tornam-se disponíveis para o próximo processo, D233;

(4) Executa o processo D233
- Recebe CW_1B e CW_2B;
- Calcula as $(\hat{CW}_I's)_{am}$ (interpolação – CW's com a mesma dimensão);
- As $(\hat{CW}_I's)_{am}$ tornam-se disponíveis para D234.

(5) Executa o processo D235:
- Recebe $\hat{P}_{rf}(n_{sq(i-1)})$ e $\hat{P}_{rf}(n_{sqi})$;
- Calcula os valores de $\hat{P}_f(n_a)$ e $\hat{P}(n_a)$;
- Os valores de $\hat{P}_f(n_a)$ tornam-se disponíveis para D250 e os de $\hat{P}(n_a)$ para D234.

(6) Executa o processo D234:
- Recebe $\hat{P}(n_a)$ e as $(\hat{CW}_I's)_{am}$;
- Calcula as $(\hat{CW}'s)_{am}$;
- As $(\hat{CW}'s)_{am}$ tornam-se disponíveis para D260.

(7) Executa o processo D236:

Atualiza a CW_1 e a CW_2:
- CW_1 <= $\hat{CW}'_{sqi}$ (CW_atual);
- CW_2 <= $\hat{CW}'_{sq(i+1)}$ (CW_posterior);

A CW_1 e a CW_2 tornam-se disponíveis para o próximo processo, D231:
- CW_1 como (CW_anterior);
- CW_2 como (CW_atual).

(8) Fim do Módulo 3.

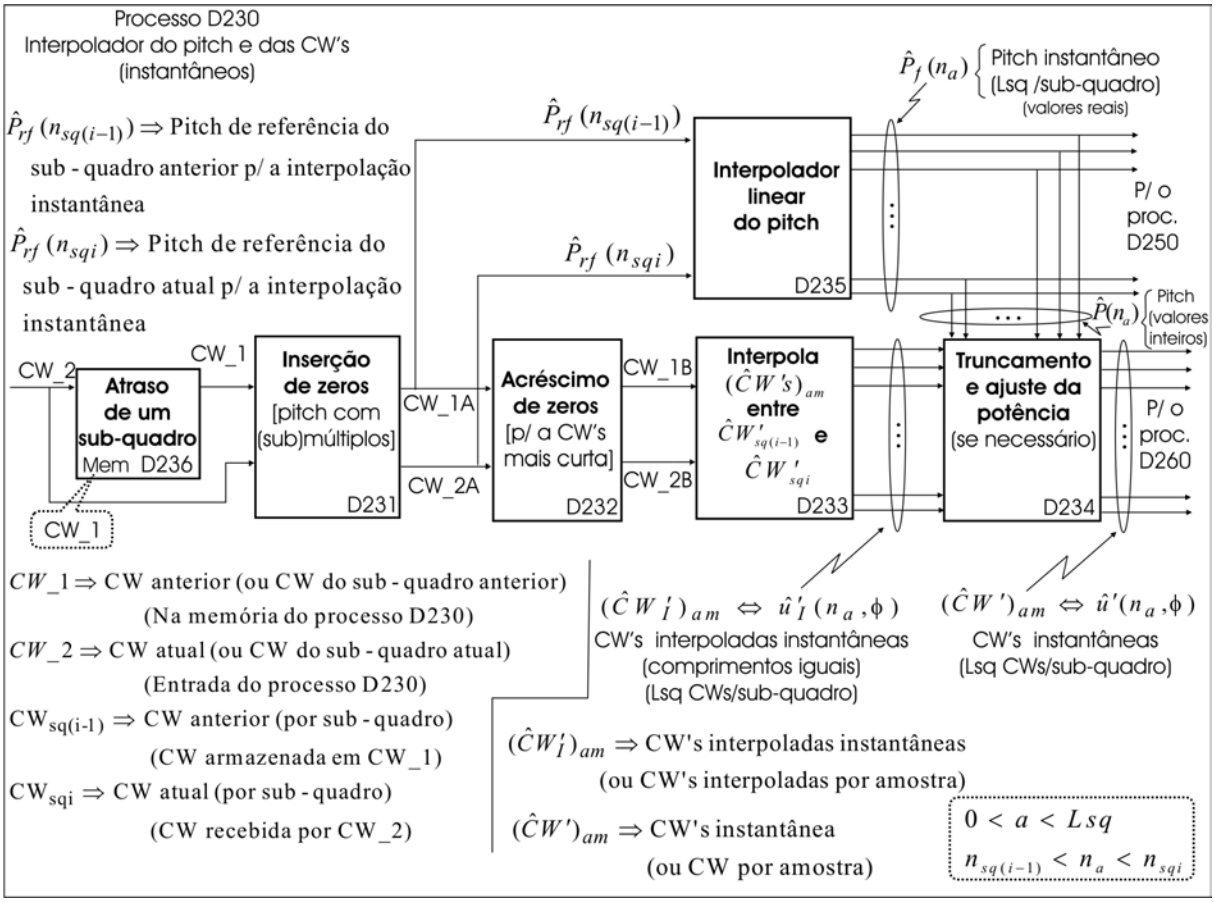

Figura 4.39 – Diagrama de blocos do (processo D230 Expandido – Interpolador do pitch e das CW's) componente do (processo D200 – bloco de Síntese - Decodificador) no codificador WI com **i**ndicação dos processos e dos parâmetros de entrada e de saída. A CW_1 (do sub-quadro anterior) é mantida na memória (Mem) enquanto a *entrada* do processo D230 recebe a CW_2 (do sub-quadro atual) resultando na *saída* as CW's e os pitch interpolados por amostra.

4.3.4.5 Estimação do Caminho da Fase

Nesta etapa, para a estimação do caminho da fase, a superfície $\hat{u}'(n_a,\phi)$ está pronta com as $(\hat{C}W's)_{am}$ do sub-quadro atual. O valor do pitch instantâneo $\hat{P}_f(n_a)$ está calculado e falta portanto calcular a fase $\phi(n_a)$ para cada posição n_a de cada $(\hat{C}W')_{am}$ para o cálculo posterior da amostra do sinal residual em n_a, $\hat{e}(n_a)$. O cálculo da seqüência $\phi(n)$ que é denominada de caminho da fase permite portanto o cálculo da seqüência de amostras do sinal residual $\hat{e}(n)$. O caminho da fase no tempo contínuo n' é definido pela Equação [14, 15]:

$$\phi(n') = \phi(n'_0) + \int_{n'_0}^{n'} \frac{2\pi}{\hat{P}_f(n')}\, dn' \qquad (4.38)$$

onde $n' = n'_0$ representa o instante do tempo inicial. Considerando o intervalo $[n'_0 = n_{(a-1)}] \le n' \le [n' = n_a]$ e supondo que o pitch evolui linearmente neste intervalo, então a Equação (4.38) pode ser aproximada por [14]:

$$\phi(n_a) \cong \phi(n_{(a-1)}) + \pi\left(\frac{1}{\hat{P}_f(n_{(a-1)})} + \frac{1}{\hat{P}_f(n_a)}\right) \tag{4.39}$$

onde n_a é o instante de amostragem atual e n_{a-1} é o instante de amostragem anterior. A seleção do valor inicial da fase $\phi(n'_0)$ não afeta a qualidade perceptível do sinal da fala [14], e portanto $\phi(n'_0)$ pode assumir qualquer valor. A Figura 4.40 mostra o esquema do processo D250.

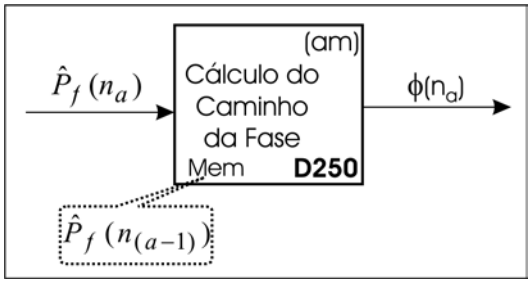

Figura 4.40 – Diagrama de blocos do (processo D250 – Cálculo do Caminho da Fase componente do (processo D200 – bloco de Síntese - Decodificador) no codificador WI com a indicação dos parâmetros de entrada e de saída. O valor $\hat{P}_f(n_{(a-1)})$ é mantido na memória (Mem) enquanto na *entrada* o processo D250 recebe o valor do pitch instantâneo $\hat{P}_f(n_a)$ (ou pitch interpolado por amostra) que resulta na *saída* a fase instantânea $\phi(n_a)$, (ou fase por amostra).

Assim o processo D250 recebe para cada posição n_a, o valor $\hat{P}_f(n_a)$ e com o valor $\hat{P}_f(n_{(a-1)})$ armazenado na memória aplica-se a Equação (4.39), calculando a correspondente fase $\phi(n_a)$. O valor de fase é limitado à faixa $0 \le \phi(n_a) < 2\pi$. Quando $\phi(n_a)$ extrapola esta faixa, faz-se então a redução para a faixa $[0, 2\pi)$. Os valores de $\phi(n_a)$ para o sub-quadro atual tornam-se disponíveis para o próximo processo D260 – Conversão 2D para 1D.

4.3.4.6 Transformação do sinal em 2D para 1D (conversão para sinal residual)

Como foi descrito no ítem 4.3.4.4 deste capítulo, a superfície bidimensional $\hat{u}'(n_a, \phi)$ possibilita a reconstrução do sinal residual. Nesta fase a superfície bidimensional está pronta

com todos os parâmetros necessários para o cálculo do sinal residual utilizando-se o processo D260 (*Conversão 2D para 1D*). Para cada posição n_a ao longo do eixo de evolução das formas de ondas existe uma forma de onda característica, $(\hat{C}W')_{am} \Leftrightarrow \hat{u}'(n_a, \phi(n))$, representada pelos coeficientes de Fourier $\{\hat{A}'_k(n_a), \hat{B}'_k(n_a)\}$, o pitch, $\hat{P}(n_a)$, e a fase $\phi(n_a)$. A conversão é simples e direta. Ela pode ser realizada calculando o valor da amplitude na superfície para o tempo $n = n_a$ na fase $\phi(n_a)$. Assim $\hat{e}[n_a] = \hat{u}'(n_a, \phi(n_a))$ onde $\hat{e}[n_a]$ é a amostra reconstruída do sinal residual e $\phi(n_a)$ é o valor da fase, na forma de onda característica $(\hat{C}W')_{am}$, correspondente à posição n_a. O cálculo de $\hat{e}[n] = \hat{u}'(n, \phi(n))$ é realizado, amostra por amostra, utilizando-se a Equação (4.6) reescrita aqui para $n = n_a$

$$\hat{e}[n_a] = \hat{u}'(n_a, \phi(n_a)) = \sum_{k=1}^{\lfloor \hat{P}(n_a)/2 \rfloor} [\hat{A}'_k(n_a)\cos(k\phi(n_a)) + \hat{B}'_k(n_a)\,\mathrm{sen}(k\phi(n_a))], \quad 0 \le \phi(.) < 2\pi,$$

$$\begin{cases} com \quad \lfloor \hat{P}(n_a)/2 \rfloor \Rightarrow \hat{P}(n_a)/2, \quad para\ \hat{P}\,(par), \\ e \quad \lfloor \hat{P}(n_a)/2 \rfloor \Rightarrow (\hat{P}(n_a)-1)/2, \quad para\ \hat{P}\,(ímpar) \end{cases} \tag{4.40}$$

onde $(0 < a < Lsq)$ e $(n_{sq(i-1)} < n_a < n_{sq(i-1)})$ para o sub-quadro atual. Esta operação está esquematizada genericamente na Figura 4.2 e em detalhes na Figura 4.41.

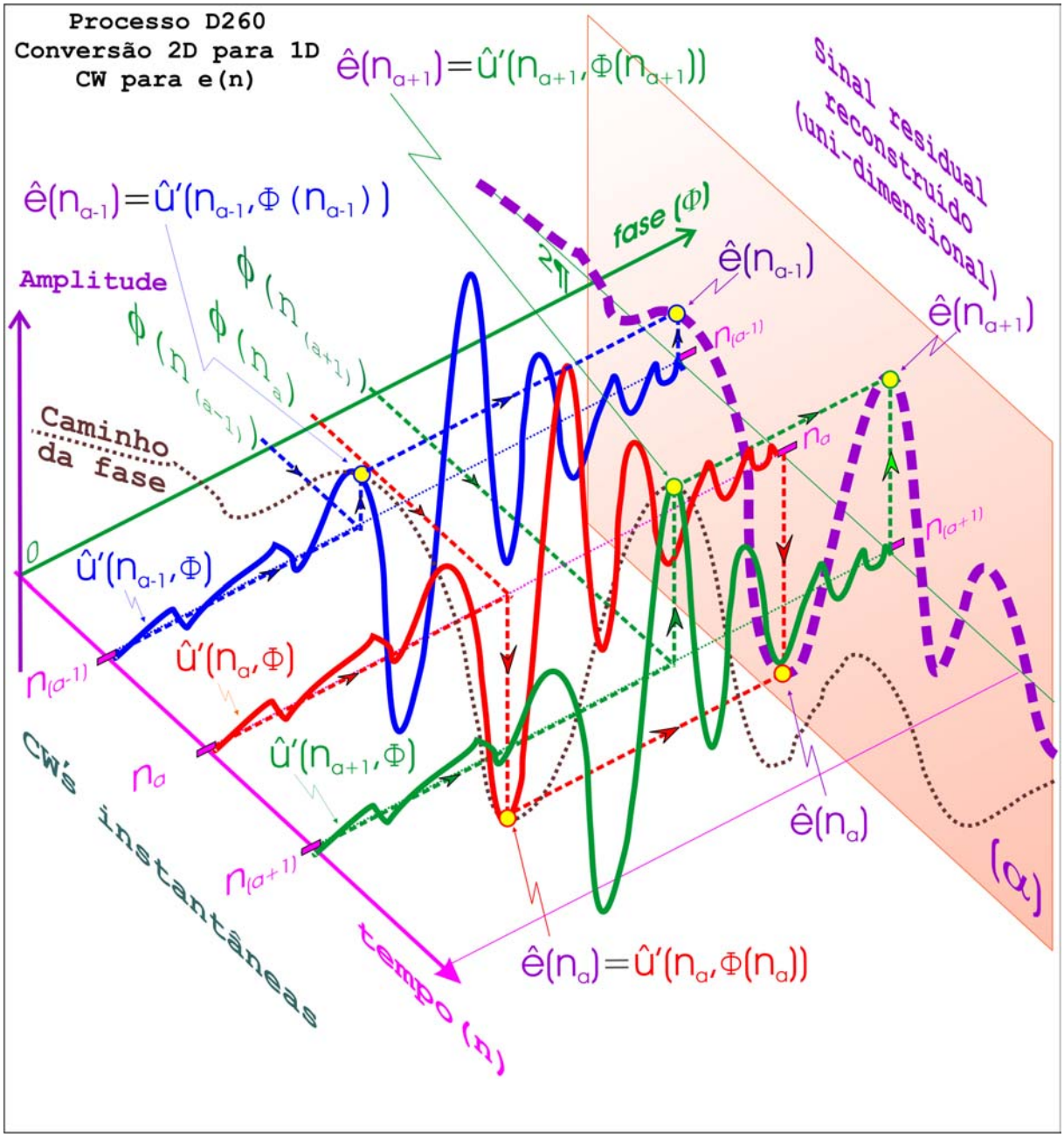

Figura 4.41 – Diagrama esquemático das operações realizadas no (processo D260 – conversão 2D para 1D) componente do (processo D200 – bloco de síntese - decodificador) no codificador WI. O diagrama mostra três $(\hat{C}W')_{am} \Leftrightarrow \hat{u}'(n_a, \phi(n))$ nos instantes n_{a-1}, n_a e n_{a+1}, com indicação das fases $\phi(n_{a-1})$, $\phi(n_a)$ e $\phi(n_{a+1})$ com as correspondentes amplitudes $\hat{e}(n_{a-1})$, $\hat{e}(n_a)$ e $\hat{e}(n_{a+1})$ para as amostras do sinal residual reconstruído, projetadas no plano (α) perpendicular ao eixo (ϕ) e paralelo ao eixo (n).

A Figura 4.42 mostra o diagrama esquemático para o processo D260 - *Conversão 2D para 1D*.

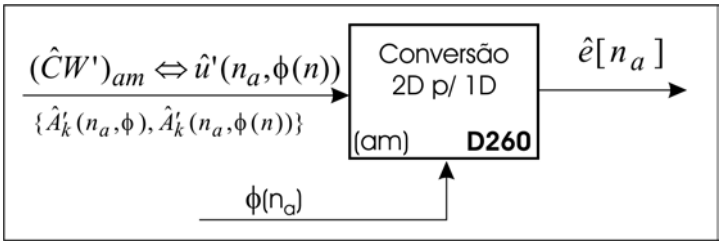

Figura 4.42 – Diagrama de blocos do (processo D260 – conversão 2D para 1D) componente do (processo D200 – bloco de síntese-decodificador) no codificador WI com indicação dos parâmetros de entrada e de saída. Na entrada recebe as $(\hat{C}W')_{am} \Leftrightarrow \hat{u}'(n_a,\phi(n))$ e a fase instantânea $\phi(n_a)$ (ou fase por amostra) que resultam na saída, o sinal residual, $\hat{e}[n_a]$ (por amostra).

Assim o (processo D260 – conversão 2D para 1D) recebe para cada posição n_a: - a $(\hat{C}W')_{am} \Leftrightarrow \hat{u}'(n_a,\phi(n))$; - $\hat{P}(n_a)$; e a fase $\phi(n_a)$. Aplica-se a Equação (4.40) e calcula-se a amostra $\hat{e}[n_a]$ do sinal residual reconstruído. Esta operação é realizada em cada sub-quadro atual $(\hat{e}[n_a])_{sqi}$ sucessivamente até completar o quadro atual $(\hat{e}[n]_l)$. As amostras do sinal residual por sub-quadro $(\hat{e}[n])_{sqi}$ tornam-se então disponíveis para o processo D280 onde é realizada a síntese do sinal da fala.

4.3.4.7 Síntese com predição linear (síntese LP)

O sinal da fala sintetizado $(\hat{s}[n])$ é obtido neste último processo D280 – *Filtro de Síntese LP*, onde o sinal residual reconstruído $(\hat{e}[n])$ é utilizado para excitar o filtro de síntese LP. Esta é a operação inversa à operação descrita no item *Análise com predição linear (análise LP)*, no processo C110 – *Filtro de análise LP*. A Figura 4.43 mostra o diagrama esquemático para o filtro de síntese LP.

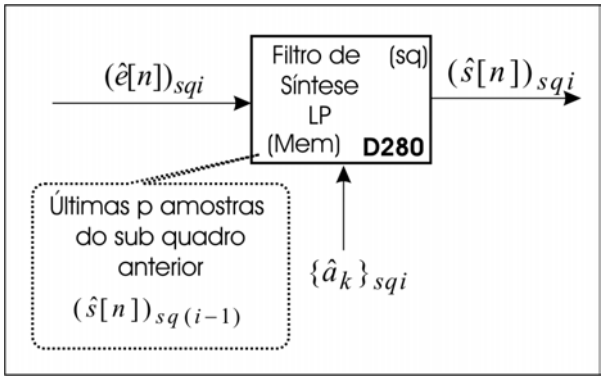

Figura 4.43 – Diagrama de blocos do (processo D280 – *Filtro de Síntese LP* componente do (processo D200 – bloco de síntese-decodificador) no codificador WI com indicação dos parâmetros de entrada e de saída. Na entrada recebe as amostras do sinal reconstruído (por sub-quadro) $(\hat{e}[n])_{sqi}$ e os coeficientes LPC $\{\hat{a}_k\}_{sqi}$ (por sub-quadro) que resultam na saída no sinal sintetizado da fala $(\hat{s}[n])_{sqi}$ (por sub-quadro). Na memória (Mem) do filtro ficam armazenadas as últimas p amostras do sinal sintetizado do sub-quadro anterior para o processamento do sub-quadro atual.

O filtro de síntese foi definido nas Equações (2.5) (2.10) e (2.11) sendo reescritas aqui como:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} \hat{a}_k z^{-k}} = \frac{\hat{S}(z)}{\hat{E}(z)} \tag{4.41}$$

onde $\{\hat{a}_k\}$ são os coeficientes LPC recebidos do codificador.

Assim a fala reconstruída pode ser obtida usando-se a Equação (2.49) reescrita aqui como

$$\hat{s}[n] = \hat{e}[n] + \sum_{k=1}^{p} \hat{a}_k \hat{s}[n-k]\,. \tag{4.42}$$

No bloco *Filtro de síntese LP* (processo D280), o processamento é composto pelas operações esquematizadas na Figura 4.44 para o cálculo do sinal sintetizado da fala $(\tilde{s}[n]_l)$ para o quadro atual. Inicialmente, o *Filtro de síntese LP* recebe $(\hat{s}[n])_{mem}$, as últimas p amostras do sinal da fala sintetizado do quadro anterior para compor a memória inicial do filtro, e $(\hat{e}[n]_l)$, as (Lq = 160) amostras do sinal residual reconstruído, relativas ao quadro atual que serão processadas em conjuntos consecutivos com Lsq = 20 amostras, ou seja, em sub-quadros com (Lsq = 20 amostras). A seqüência de amostras nos sub-quadros são denominadas por $(\hat{e}[n])_{sq0}, (\hat{e}[n])_{sq1}$, …, $(\hat{e}[n])_{sq7}$. O *Filtro de síntese LP* recebe também os coeficientes LPC (por sub-quadro), $\{\hat{a}_k\}_{sq0}$, $\{\hat{a}_k\}_{sq1}$,…, $\{\hat{a}_k\}_{sq7}$ (onde $1 \le k \le p$) os quais são convertidos

para os coeficientes do filtro do erro de predição $\{\hat{ec}_k\}_{sq0}, \{\hat{ec}_k\}_{sq1}, \ldots, \{\hat{ec}_k\}_{sq7}$, onde $0 \le k \le p$. Para cada sub-quadro, $(\hat{e}[n])_{sqi}$, com o seu respectivo vetor de coeficientes do filtro do erro de predição, $\{\hat{ec}_k\}_{sqi}$ (por sub-quadro), executa-se a filtragem (a síntese), utilizando-se a equação (4.42), obtendo-se, $(\hat{s}[n])_{sqi}$, o correspondente sub-quadro do sinal da fala sintetizado (e enfatizado). Após o cálculo, todos os sub-quadros $(\hat{s}[n])_{sq0}$, $(\hat{s}[n])_{sq1}$, …, $(\hat{s}[n])_{sq7}$ que compõem o sinal sintetizado enfatizado para o quadro atual $(\hat{s}[n]_l)$ são então dê-enfatizados, aplicando-se a Equação (2.78) com o mesmo valor para o fator, $\alpha = 0.9$ $(\lambda = \alpha)$, utilizado durante o processo de pré-ênfase. Após este processo os sub-quadros $(\tilde{s}[n])_{sq0}$, $(\tilde{s}[n])_{sq1}$, …, $(\tilde{s}[n])_{sq7}$ que compõem o quadro atual do sinal sintetizado da fala, $(\tilde{s}[n]_l)$, tornam-se disponíveis para a reprodução e/ou o armazenamento.

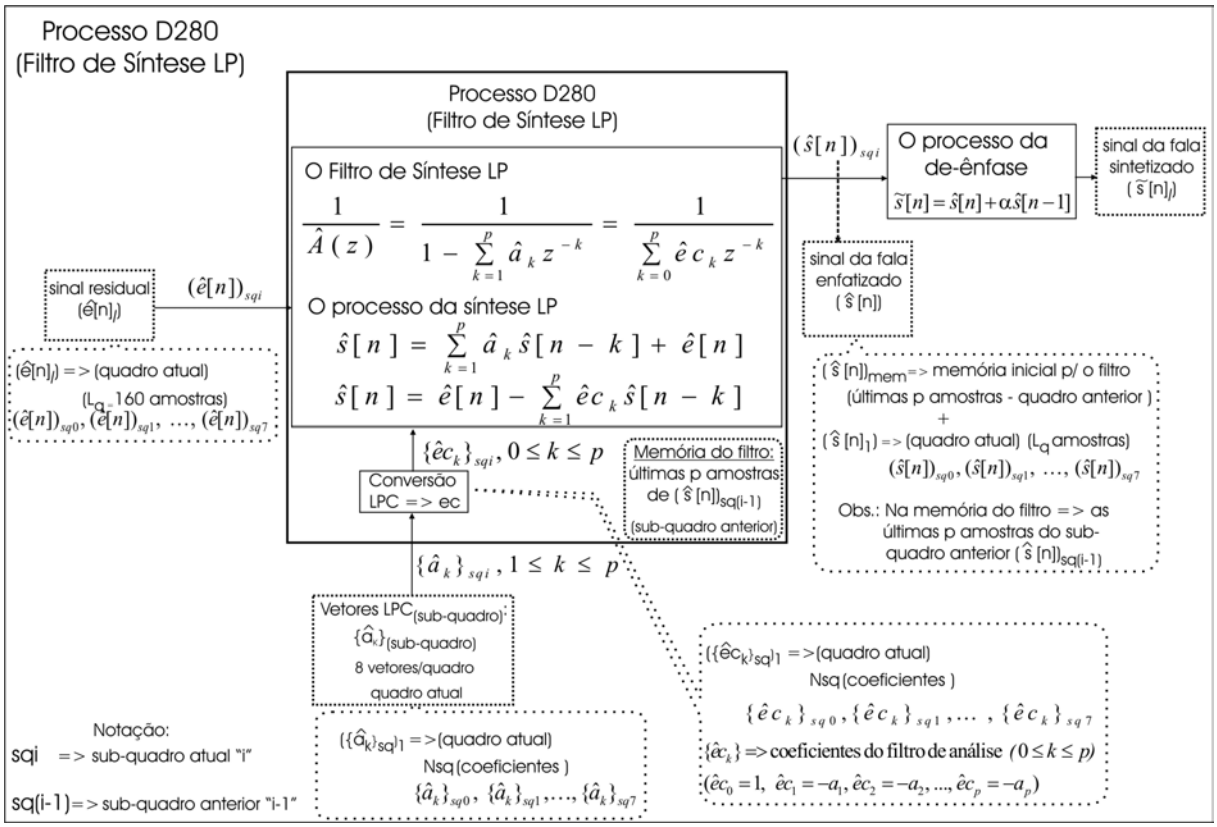

Figura 4.44 – Diagrama esquemático das operações no bloco filtro de síntese LP (processo D280) do codificador WI – Equações e notação dos parâmetros.

4.3.4.8 O Diagrama de blocos com indicação dos parâmetros - processo D200 (bloco de síntese)

Na Figura 4.45 é apresentado o diagrama de blocos de síntese expandido (processo D200) do codificador WI com indicação dos parâmetros de entrada e de saída nos blocos que foram mostrados na Figura 4.34.

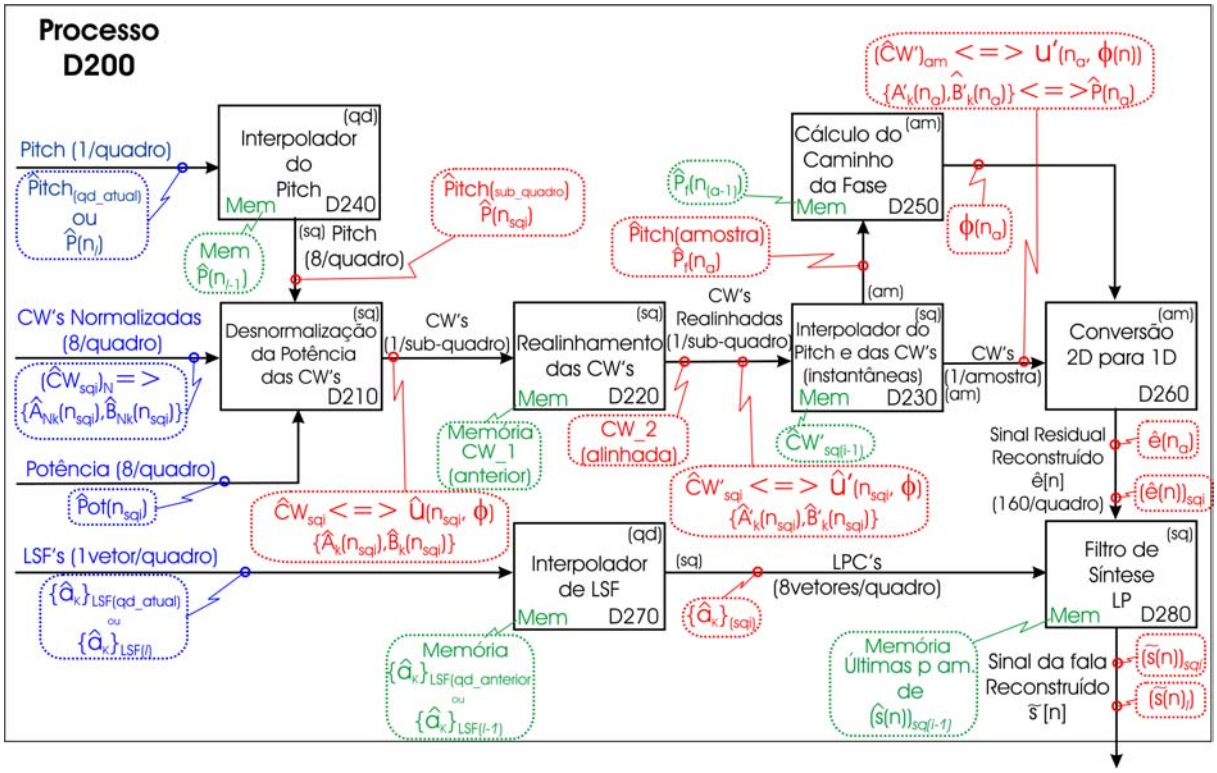

Figura 4.45 – Diagrama de blocos de síntese expandido (processo D200) do codificador WI com indicação dos parâmetros de entrada e de saída nos blocos.

4.4 Considerações finais sobre o capítulo

Neste capítulo foi apresentada uma visão geral da técnica de codificação da fala por interpolação de ondas, técnica WI, e a estrutura básica para um codificador/decodificador utilizando a técnica WI, sem o uso da compressão dos parâmetros.

O objetivo principal foi descrever com detalhes os processos relativos à camada de análise-síntese envolvidos no sistema codificador/decodificador para facilitar o entendimento e permitir a visualização de um algoritmo para a implementação de um simulador para o sistema codificador/decodificador WI na camada análise-síntese. Foram apresentados detalhes dos processos e dos procedimentos envolvidos, e alguns algoritmos os quais não estão disponíveis na literatura.

A partir deste capítulo, e auxiliado pelas informações em [14], foi possível implementar, em linguagem C, um simulador para o codificador WI na camada de análise-síntese, que se tornou a parte principal deste trabalho. A visualização e a implementação do algoritmo para o codificador/decodificador WI foram necessárias, pois o algoritmo não está disponível na literatura científica devido a interesses comerciais. Portanto, fica caracterizado que a partir das informações deste trabalho e das informações em [14] é possível implementar um codificador/decodificador WI básico e convencional, ou seja, um sistema de análise-síntese WI.

Capítulo 5

RESULTADOS DO SISTEMA DE ANÁLISE-SÍNTESE WI

5.1 Introdução

Neste capítulo são apresentados os resultados das simulações que mostram o desempenho do sistema de análise - síntese WI convencional e básico ou *sistema de análise – síntese WI (padrão)*. Também são apresentados os resultados preliminares para um sistema de análise – síntese WI com melhoria no processo de extração das CW's no lado da análise, através de interpolação das CW's nas posições regulares. Para simplificação este sistema será denominado de *sistema de análise – síntese WI (com interpolação)*.

Nas simulações foram utilizados sinais da fala amostrados com uma freqüência de 11025 Hz a partir da fala de um locutor – adulto, de uma locutora – adulta e de uma locutora – infantil os quais pronunciaram palavras do português falado no Brasil que foram gravadas em arquivos no formato wave, *nomearq*.wav – extensão ".wav". Na gravação dos arquivos foi utilizado um sistema de aquisição de dados da CREATIVE LAB – placa Sound Blaster Live, com conversores A/D e D/A de 16 bits, instalado em um PC com um processador AMD K6 III de 450 MHz.

Os *sinais da fala originais* em arquivos "*nomearq*.wav" foram processados nos simuladores, sistemas de análise - síntese WI, a partir dos quais foram obtidos os seguintes tipos de arquivos: (a) no analisador: (após a análise) - arquivo com os sinais residuais, considerados como *sinais residuais originais (nomearq*_res.wav*)*; (b) no sintetizador: (antes da síntese) - arquivo com os *sinais residuais reconstruídos (nomearq*_res_rec.wav*)*; e (após a síntese) - o arquivo com os *sinais da fala sintetizados* (ou reconstruídos) *(nomearq*_sin_rec.wav*)*. O esquema do sistema com os tipos de arquivos são mostrados na Figura 5.3.

5.2 Os resultados

Para mostrar o desempenho dos sistemas de análise - síntese WI são apresentados: (a) os sinais da fala originais e os sinais da fala sintetizados; e (b) os sinais residuais originais, após a análise, e os sinais residuais reconstruídos, no lado do sintetizador.

Como resultados são apresentados: (a) os gráficos das formas de onda (amplitude *versus* tempo) de alguns sinais da fala e sinais residuais da fala para uma comparação visual

(avaliação subjetiva); (b) a medida PESQ_MOS obtida a partir da aplicação do método PESQ [5] como uma avaliação objetiva da qualidade perceptual; e (c) a medida SNRSEG-NCCF obtida a partir da aplicação do método da Relação Sinal Ruído Segmental (SNRSEG) com o auxílio da "Normalized Cross Correlation Function" (NCCF). Este método, SNRSEG – NCCF, é aplicado aos sinais da fala e aos sinais residuais como uma avaliação objetiva da reconstrução dos sinais e na avaliação da defasagem entre os sinais originais e reconstruídos.

5.2.1 As medidas para avaliação do desempenho dos sistemas de análise – síntese WI

O sistema de análise – síntese WI gera um sinal reconstruído da fala sem sincronismo temporal com o sinal original da fala [14]. Assim a forma mais adequada de avaliar o sistema seria através de métodos subjetivos de avaliação da qualidade perceptual da fala, ou seja, por meio de testes auditivos com vários ouvintes. O método PESQ [5] foi desenvolvido para simplificar esta avaliação substituindo os testes auditivos, que são longos e trabalhosos. Assim, neste trabalho o método PESQ é utilizado para fazer uma avaliação objetiva da qualidade perceptual dos sinais da fala reconstruídos pelos sistemas. Também devido à falta de sincronismo a relação sinal ruído (SNR) não pode ser aplicada diretamente para avaliar o sinal reconstruído de forma objetiva [14]. Uma alternativa para fazer uma avaliação objetiva foi aplicar a relação sinal ruído segmental (SNRSEG) auxiliada pela função de correlação cruzada normalizada (NCCF) procurando sincronizar cada segmento do sinal original com as amostras mais similares na região próxima ao segmento correspondente no sinal reconstruído da fala. O método SNRSEG – NCCF é utilizado também para fazer uma avaliação da defasagem (ou falta de sincronismo) entre os sinais originais da fala e os sinais reconstruídos e entre os sinais residuais da fala no lado da análise (analisador) e sinais residuais reconstruídos da fala no lado da síntese (sintetizador).

A seguir são apresentados os métodos: "Perceptual Evaluation of Speech Quality" (PESQ), denominado de método PESQ; e o método da Relação Sinal Ruído Segmental (SNRSEG) com o auxílio da "Normalized Cross Correlation Function" (NCCF), denominado de método da SNRSEG –NCCF .

5.2.1.1 O método PESQ - a medida PESQ_MOS é obtida a partir da aplicação do método PESQ como uma avaliação objetiva da qualidade perceptual [3, 4, 5, 6]. Este método foi introduzido na seção 1.3.3.3, Medidas objetivas de predição da qualidade subjetiva da fala, do capítulo 1 deste trabalho. São usados durante a avaliação PESQ_MOS: o sinal original da fala (como padrão de referência) e o sinal de saída, ou sinal da fala reconstruído (como sinal em teste) após sua transição pelo sistema de análise - síntese da fala em teste; ou, o sinal residual

da fala durante a análise (como padrão de referência) e o sinal residual da fala reconstruído antes da síntese, como sinal em teste.

O método faz o processamento do sinal da fala considerando as medidas perceptuais. Ele é estruturado em três etapas: pré-processamento [12, 13], modelamento psico-acústico, e modelamento cognitivo como mostrado na Figura 5.1. No pré-processamento ambos os sinais são equalizados para um nível de potência padrão equivalente ao nível acústico que é usado

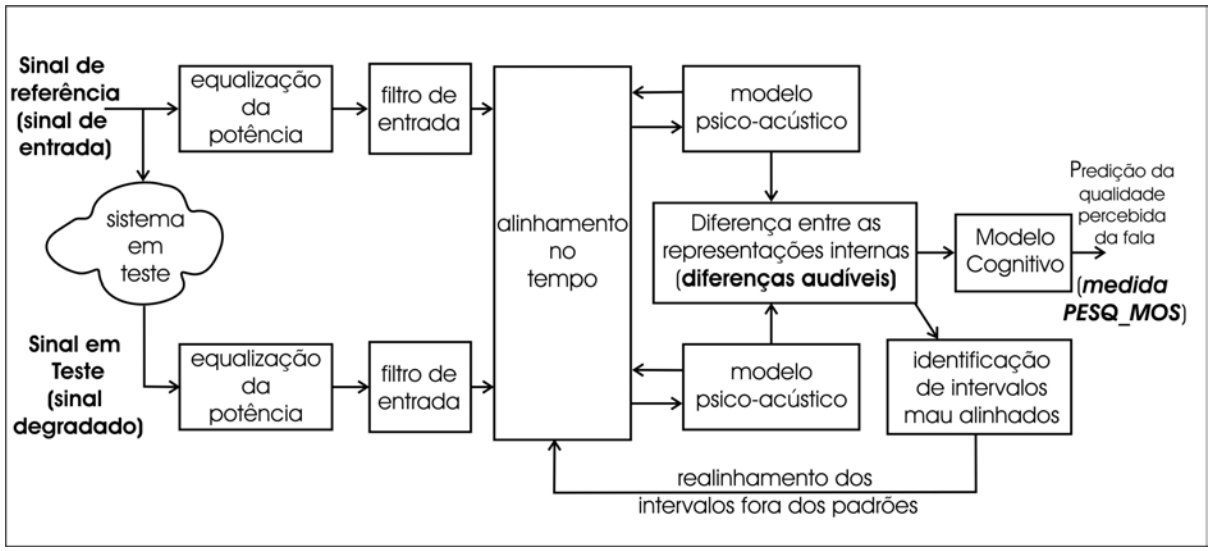

Figura 5.1 – Diagrama esquemático das etapas do método PESQ (adaptado de [5]).

normalmente em testes subjetivos. Os sinais são filtrados usando Fast Fourier Transform (FFT) com um filtro de entrada para simular o padrão de um canal de comunicação telefônica e em seguida são alinhados no tempo. Os sinais são então processados por um modelamento psico-acústico, que imita as propriedades principais do sistema de audição humano, relacionando os sinais com uma representação do nível acústico percebido no domínio do tempo e da freqüência. A partir do modelamento psico-acústico são extraídos os parâmetros relacionados com as diferenças entre os dois sinais. Estes parâmetros, no tempo e na freqüência, são então avaliados pelo modelamento cognitivo que os relacionam no tempo e na freqüência para predizer a medida subjetiva MOS através da medida PESQ_MOS. O modelo cognitivo faz uma avaliação do julgamento subjetivo da qualidade da fala. Esta avaliação é realizada pela determinação da distância perceptual entre o sinal medido e o sinal de referência e então cria uma figura de mérito que descreve a qualidade da fala. A figura de mérito é geralmente uma função não linear do valor MOS determinado subjetivamente. Com o intuito de conseguir um estimador para o valor MOS, foi realizado o mapeamento do resultado objetivo para a escala MOS. Assim as medidas PESQ_MOS variam de – 0,5 (sinal completamente distorcido) até 4,5 (sinal sem distorção), enquanto a escala MOS, mostrada na

Tabela 1.1 no capítulo 1 deste trabalho, varia de 1 (sinal com nível de distorção muito irritante) até 5 (sinal com nível de distorção imperceptível).

Neste trabalho as estimações das medidas PESQ_MOS, são obtidas a partir da aplicação dos algoritmos da "Perceptual Evaluation of Speech Quality" (PESQ) recomendação P.862 do ITU-T [5].

5.2.1.2 O método da SNRSEG – NCCF - A medida SNRSEG-NCCF é obtida a partir do cálculo da Relação Sinal Ruído Segmental (SNRSEG) com o auxilio da "Normalized Cross Correlation Function" (NCCF). O método é aplicado para uma avaliação objetiva dos sinais reconstruídos:

(a) Sinais reconstruídos da fala:
Sinal de entrada: sinal da fala original;
Sinal degradado: sinal da fala reconstruído (síntese).

(b) Sinais reconstruídos residuais:
Sinal de entrada: sinal residual original (análise);
Sinal degradado: sinal residual reconstruído (síntese).

Os objetivos do método da SNRSEG – NCCF são: (1) Determinar a relação sinal ruído segmental SNRSEG; e (2) Verificar a partir da aplicação da NCCF, se existe defasagem entre os sinais originais e reconstruídos, se a defasagem é constante ou variável e o grau da defasagem.

Para alcançar estes objetivos o sinal original é dividido em segmentos com M amostras. O sinal degradado (ou reconstruído) também é divido em segmentos com M amostras em posições correspondentes aos segmentos do sinal original. Para cada segmento do sinal reconstruído monta-se uma janela deslizante com as M amostras do segmento acrescido com M/2 amostras à esquerda e M/2 amostras à direita do segmento. A esta janela é permitido deslizar (deslocar-se) de até M/2 amostras para a direita ou de até M/2 amostras para a esquerda na procura das amostras no sinal reconstruído em correspondência com o segmento no sinal original que tenham o maior valor para a NCCF, ou seja, na procura do segmento com maior similaridade. Assim para cada segmento calcula-se: o deslocamento, *d',* do segmento no sinal reconstruído correspondente ao valor máximo da NCCF; o valor máximo da NCCF, $R_{máx}$; e a relação sinal ruído, *SNR*, entre os segmentos com maior similaridade. E globalmente para o sinal reconstruído são calculados: a média dos deslocamentos *d's,* $d'_{méd}$, para os segmentos onde ocorre o máximo da NCCF; o valor médio das diferenças absolutas dos deslocamentos, $d'_{méd\_dif\_abs}$; e a medida *SNRSEG – NCCF*. A Figura 5.2 mostra o esquema para aplicação do método da SNRSEG – NCCF.

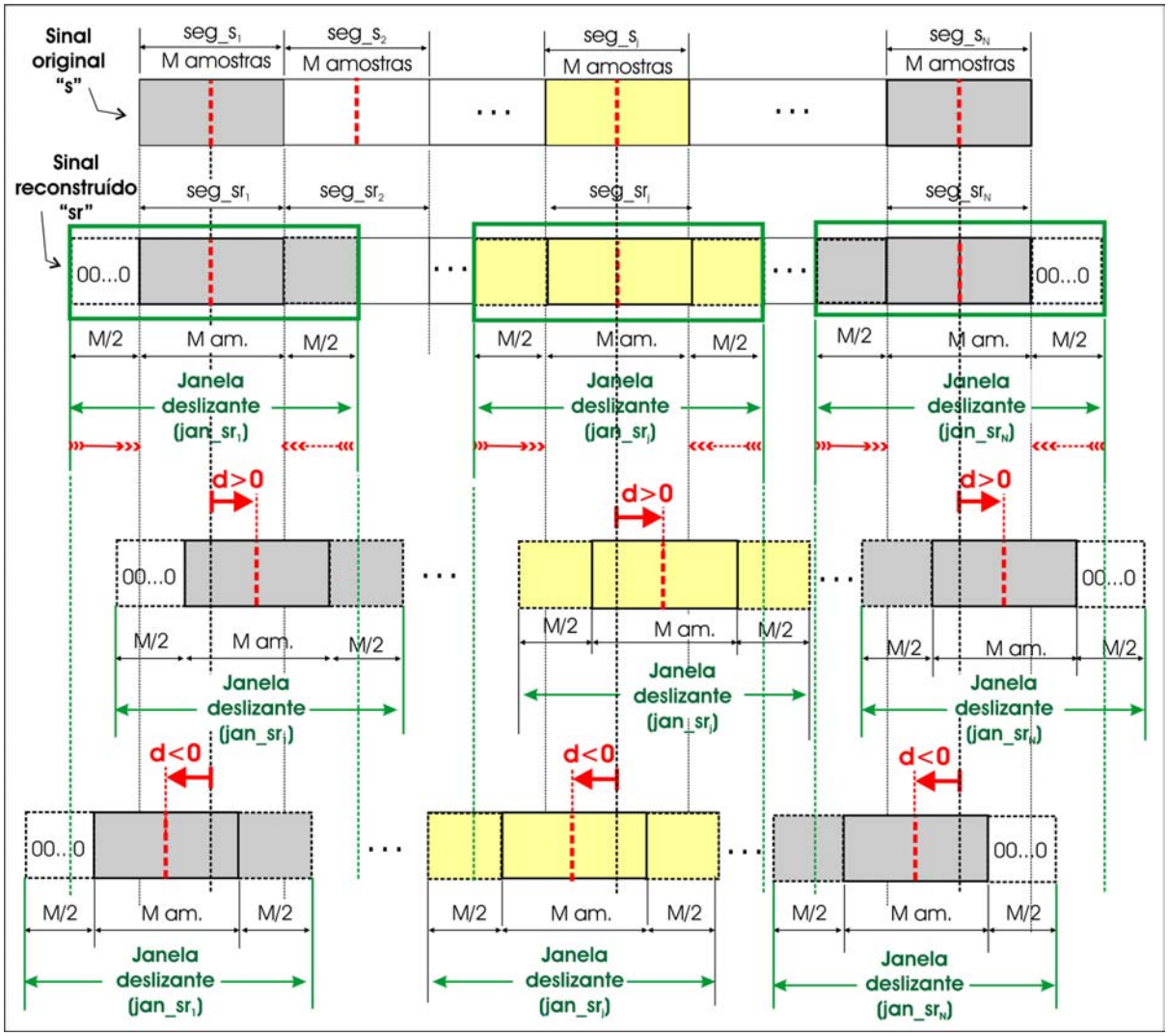

Figura 5.2 – Esquema para aplicação do método relação sinal ruído segmental usando a função da correlação cruzada normalizada (SNRSEG – NCCF).

O algoritmo para o cálculo da SNRSEG – NCCF:

(1) O sinal de entrada (sinal original, s) é segmentado em N segmentos com M amostras, ou seja, (seg_s_1, seg_s_2, ..., seg_s_j, ..., seg_s_N);

(2) O sinal degradado (sinal reconstruído, sr) também é segmentado da mesma forma, em posições correspondentes às posições do sinal original, em (seg_sr_1, seg_sr_2, ..., seg_sr_j, ..., seg_sr_N);

(3) Em torno do j-ésimo segmento em processamento, seg_sr_j, com M amostras, para o sinal degradado, define-se uma janela deslizante na memória, jan_sr_j, com M/2 posições à esquerda e M/2 posições à direita das M posições do segmento. A janela deslizante jan_sr_j, fica então com (M/2 + M + M/2) posições na memória para as amostras do sinal degradado;

(3.a) *Para o primeiro segmento* (seg_sr_1), a janela deslizante correspondente jan_sr_1, é formada com M/2 posições com valores iguais a zero + M amostras do segmento em correspondência às M amostras do sinal original + M/2 amostras à direita do segmento;

(3.b) *Para os segmentos seguintes* (seg_2, seg_3, ... , seg_sr_{N-1}) as janelas deslizantes ficam completas com (M/2 + M + M/2) amostras do sinal degradado;

(3.c) *Para o último segmento* (seg_sr_N), a janela jan_sr_N, é formada com M/2 amostras à esquerda do segmento + M amostras do segmento (seg_sr_N) + M/2 posições à direita do segmento com valores iguais a zero;

(4) À janela deslizante jan_sr_j é permitido um deslocamento d, em relação ao centro do segmento seg_s_j (com M amostras) na faixa $-d_{máx} < d < d_{máx}$, onde $d_{máx}$ = M/2 amostras;

(4.a) Para cada posição *d* da janela jan_sr_j, calcula-se o valor da NCCF, *R*, das M amostras em posições coincidentes com as posições das M amostras do segmento do sinal original seg_s_j;

(4.b) A posição da janela, *d'*, que corresponde ao valor máximo da NCCF, $R_{máx}$, determina na janela jan_sr_j, o segmento no sinal degradado com maior similaridade ao segmento seg_s_j que, na seqüência do processamento, participa do cálculo da *SNRSEG*; Armazena-se a posição *d',* a somatória do valor absoluto da diferença (*d'atual - d'anterior)* e o valor máximo da NCCF, $R_{máx}$;

(4.c) Calcula-se, utilizando a Equação 1.2, a *SNR* entre o segmento do sinal original e o segmento do sinal degradado com maior similaridade; Armazena-se a *SNR* para o segmento seg_s_j;

(4.d) Repete-se o procedimento para todos os segmentos;

(5) Calcula-se a *SNRSEG-NCCF* utilizando-se a Equação 1.3;

(6) Calcula-se o deslocamento médio, $d'_{méd}$, a partir dos valores dos d's usando

$$d'_{méd} = \frac{\sum_{i=1}^{N} d'_i}{N}; \qquad (5.1)$$

(7) Calcula-se o valor médio das diferenças absolutas dos deslocamentos, $d'_{méd\_dif\_abs}$, a partir da somatória do valor absoluto da diferença (*d'atual - d'anterior)* usando

$$d'_{méd\_dif\_abs} = \frac{\sum_{i=1}^{N} |d'_i - d'_{i-1}|}{N} \qquad (5.2)$$

onde $d'_0 = 0$;

(8) Tornam-se disponíveis para avaliação:

Para cada segmento: (a) o deslocamento *d'*; (b) o valor máximo da NCCF, $R_{máx}$; e (c) a relação sinal ruído, *SNR*.

E globalmente para o sinal reconstruído: (a) o valor médio dos deslocamentos d's, $d'_{méd}$; (b) o valor médio das diferenças absolutas dos deslocamentos, $d'_{méd\_dif\_abs}$; e (c) a medida da *SNRSEG – NCCF*.

(9) Fim.

5.2.2 Os sinais da fala utilizados

Os sinais da fala originais foram gravados nos arquivos:

(1) *casa1*.wav e *ebonita1*.wav por uma locutora adulta;

(2) *casa2*.wav e *ebonita2*.wav por um locutor adulto;

(3) e *casa3*.wav e *ebonita3*.wav por uma locutora infantil (4 anos de idade);

5.2.3 O esquema para as simulações dos sistemas de análise – síntese WI

Para cada expressão de fala pronunciada pelos locutores foram realizadas simulações no *sistema de análise – síntese WI (padrão)* e no *sistema de análise – síntese WI (com interpolação)*. Durante as simulações os valores para o pitch, P, foram estimados na faixa, $P_{mín} \leq P \leq P_{máx}$ com $P_{mín}$ e $P_{máx}$ fixos respectivamente em 20 e 120 amostras.

Em cada simulação os sinais da fala originais, gravados em arquivos tipo *nomearq*.wav, são processados primeiramente no analisador onde são obtidos os sinais residuais, gravados em arquivo tipo *nomearq*_res.wav, e os parâmetros do sistema: o pitch, as CW's (coeficientes de Fourier), a potência das CW's e os coeficientes LSF's que são armazenados em um arquivo tipo *par_wi.c*. O sintetizador recebe os parâmetros através do arquivo *par_wi.c* que são processados resultando no sinal residual reconstruído da fala, gravados em arquivo tipo *nomearq*_res_rec.wav, e no sinal da fala sintetizado (ou reconstruído), gravados em arquivo tipo *nomearq*_sin_rec.wav. A Figura 5.3 mostra o esquema com os tipos de arquivos utilizados na simulação do sistema de análise – síntese WI deste trabalho. Aos nomes dos arquivos processados pelos sistemas de análise – síntese WI são acrescentadas as abreviaturas "pd" para o sistema de análise – síntese (padrão) e "c_interp" para o sistema de análise – síntese WI (com interpolação) como é mostrado na Figura 5.3.

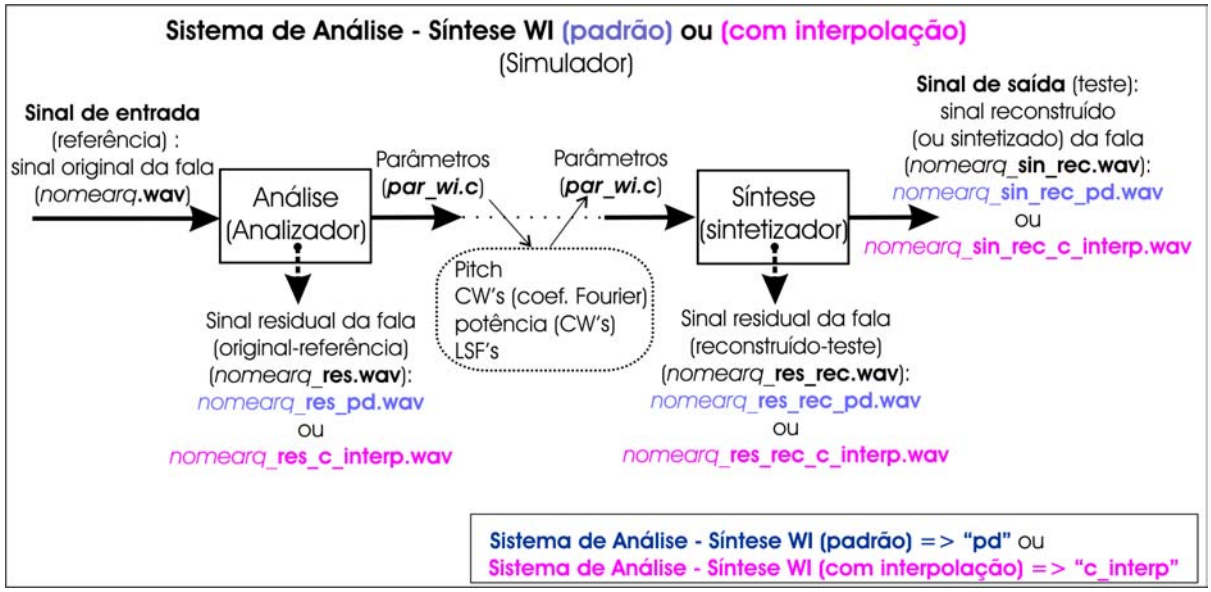

Figura 5.3 – Diagrama esquemático do simulador para o sistema de análise – síntese WI.

5.2.4 Avaliação do sistema de análise - síntese WI (padrão)

Nesta seção são apresentados os resultados para o sistema de análise – síntese WI convencional e básico ou *sistema de análise – síntese WI (padrão)*: as medidas PESQ_MOS e SNRSEG – NCCF; a defasagem dos sinais no sistema de análise – síntese WI (padrão) e os gráficos das formas de onda dos sinais no sistema.

5.2.4.1 Os resultados: medidas PESQ_MOS e SNRSEG – NCCF para o sistema de análise – síntese WI (padrão)

Nesta seção são apresentados os resultados das medidas PESQ_MOS e SNRSEG - NCCF que são mostrados nas tabelas que seguem onde são indicados os ítens:

(a) Arquivos com os sinais considerados como sinais de referência:

- o nome do arquivo com o *sinal original da fala* => ***nomearq*.wav**;
- o nome do arquivo com o *sinal residual da fala (no lado da análise)* => ***nomearq*_res_pd.wav;**

(b) Arquivos com os sinais considerados como sinais de teste (sinais reconstruídos):

- com o *sinal residual reconstruído da fala* (no lado da síntese) => ***nomearq*_res_rec_pd.wav**;
- com o *sinal da fala reconstruído (sintetizado)* => ***nomearq*_sin_rec_pd.wav**;

(c) A ***medida PESQ_MOS;***

(d) A medida ***SNRSEG - NCCF*** - a relação sinal ruído segmental com o auxílio da função de correlação cruzada normalizada. Esta medida é calculada para segmentos com tamanho M = 99 e M =128 amostras.

- O deslocamento médio, **d'med**;
- O valor médio das diferenças absolutas dos deslocamentos, ***d'***$_{méd\_dif\_abs}$.

Tabela 5.1 - Locutora – adulta - Sistema análise/síntese WI

Locutora – adulta			**Sistema análise/síntese WI (padrão)**		
Expressão da fala	**Sinal original (referência)**	**Sinal reconstruído (em teste)**	**PESQ_MOS (–0,5 a 4,5)**	**SNRSEG - NCCF (dB) [d'med/ d'med_dif_abs] (amostras) M = 99 am.**	**SNRSEG - NCCF (dB) [d'med/ d'med_dif_abs] (amostras) M = 128 am.**
"casa"	***casa1*.wav**	***casa1*_sin_rec_pd.wav**	**3,439**	**9,751488 [6/22]**	**9,29640 [8/26]**
	***casa1*_res_pd.wav**	***casa1*_res_rec_pd.wav**	**3,442**	**1,527073 [1/10]**	**1,249191 [-1/29]**
"é bonita"	***ebonita1*.wav**	***ebonita1*_sin_rec_pd.wav**	**3,372**	**14,188506 [2/5]**	**14,498303 [3/38]**
	***ebonita1*_res_pd.wav**	***ebonita1*_res_rec_pd.wav**	**3,695**	**5,254081 [-3/14]**	**4,261524 [0/23]**

Tabela 5.2 - Locutor – adulto- Sistema análise/síntese WI

Locutor – adulto			**Sistema análise/síntese WI (padrão)**		
Expressão da fala	**Sinal original (referência)**	**Sinal reconstruído (em teste)**	**PESQ_MOS (–0,5 a 4,5)**	**SNRSEG - NCCF (dB) [d'med/ d'med_dif_abs] (amostras) M = 99 am.**	**SNRSEG - NCCF (dB) [d'med/ d'med_dif_abs] (amostras) M = 128 am.**
"casa"	***casa2*.wav**	***casa2*_sin_rec_pd.wav**	**3,916**	**11,959578 [-1/6]**	**11,378476 [-2/9]**
	***casa2*_res_pd.wav**	***casa2*_res_rec_pd.wav**	**3,780**	**2,605935 [4/4]**	**2,180656 [6/6]**
"é bonita"	***ebonita2*.wav**	***ebonita2*_sin_rec_pd.wav**	**3,375**	**9,363902 [-19/11]**	**9,270845 [-12/22]**
	***ebonita2*_res_pd.wav**	***ebonita2*_res_rec_pd.wav**	**3,543**	**2,111374 [8/21]**	**2,748792 [9/33]**

Tabela 5.3 - Locutora – infantil - Sistema análise/síntese WI

Locutora – infantil			**Sistema análise/síntese WI (padrão)**		
Expressão da fala	**Sinal original (referência)**	**Sinal reconstruído (em teste)**	**PESQ_MOS (–0,5 a 4,5)**	**SNRSEG - NCCF (dB) [d'med/ d'med_dif_abs] (amostras) M = 99 am.**	**SNRSEG - NCCF (dB) [d'med/ d'med_dif_abs] (amostras) M = 128 am.**
"casa"	***casa3*.wav**	***casa3*_sin_rec_pd.wav**	**2,523**	**6,901643 [7/22]**	**6,494740 [9/33]**
	***casa3*_res_pd.wav**	***casa3*_res_rec_pd.wav**	**2,799**	**2,904033 [0/15]**	**2,519160 [-5/18]**
"é bonita"	***ebonita3*.wav**	***ebonita3*_ sin_rec_pd.wav**	**2,391**	**12,427695 [1/18]**	**12,712596 [9/33]**
	***ebonita3*_res_pd.wav**	***ebonita3*_ res_rec_pd.wav**	**2,818**	**4,103289 [-2/11]**	**3,375701 [-3/17]**

Nesta seção os resultados principais são as medidas PESQ_MOS que avaliam as qualidades perceptuais da fala e as medidas SNRSEG-NCCF que avaliam a reconstrução da forma de onda por segmentos. As medidas SNRSEG-NCCF, d'med e d'med_dif_abs também são utilizadas como dados para a avaliação da defasagem na seção 5.2.4.2, deste capítulo. Assim os resultados nas Tabelas 5.1, 5.2 e 5.3 mostram:

- *Para os sinais da fala dos locutores adultos* as medidas PESQ_MOS (com escala variando entre – 0,5 e 4,5) estão variando entre 3,372 e 3,916 o que indica (conforme a Tabela 1.1 e Figura 1.1 do capítulo 1 deste trabalho) uma fala com qualidades perceptuais entre a faixa "satisfatória" e "boa". Isto mostra que o sistema de análise – síntese WI (padrão) apresenta bom desempenho atingindo a faixa "communication quality". Com alguma melhoria, ou ajuste, o sistema de análise – síntese WI (padrão) poderá atingir a faixa "Toll quality";
- *Para os sinais da fala da locutora infantil* as medidas PESQ_MOS apresentaram resultados inferiores aos resultados para os locutores adultos. Isto está relacionado com a fala em formação (locutora com 4 anos de idade). Para o sinal da fala em ebonita3.wav e *ebonita3*_sin_rec_pd.wav a medida SNRSEG_NCCF apresentou um dos maiores valores, 12,427695 dB;
- *Para os sinais da fala* as medidas SNRSEG-NCCF apresentaram maiores valores em geral em relação às medidas SNRSEG_NCCF para *os sinais residuais*, o que indica uma melhor reconstrução das formas de onda por segmento para os sinais da fala em relação à reconstrução das formas de onda dos sinais residuais por segmento. Os sinais residuais são ruidosos com formas de onda que apresentam amostras com variação bruscas de valores enquanto os sinais da fala têm forma de onda mais suave e contínua;

- O sinal da fala em *ebonita1*_sin_rec_pd.wav apresentou a maior medida SNRSEG_NCCF = 14,498303 dB (M = 128 amostras), sendo o sinal com melhor reconstrução, enquanto o sinal em *casa2*_sin_rec_pd.wav apresentou a maior medida PESQ_MOS = 3,916, sendo o sinal com melhor qualidade perceptual;
- As tabelas mostram os resultados para o cálculo da SNRSEG-NCCF com segmentos de tamanho M igual a 99 amostras e 128 amostras. Os valores para SNRSEG-NCCF apresentaram uma pequena diferença mas indicando uma tendência de maiores valores para o cálculo com M = 99 amostras. Também para M = 99 amostras, observa-se que para todos os sinais houve uma diminuição da medida d'med_dif_abs que está relacionada com a defasagem entre os segmentos. O tamanho M = 99 amostras é um valor próximo do pitch médio para o locutor adulto.

5.2.4.2 Os resultados: A defasagem dos sinais no sistema de análise – síntese WI (padrão)

Este experimento foi realizado para avaliar a defasagem entre os sinais originais e reconstruídos. O objetivo é verificar se existe defasagem entre os sinais originais e reconstruídos; e em caso afirmativo, se ela é constante ou variável.

Na seção 5.2.1.2 deste capítulo foi descrito o método para calcular a SNRSEG – NCCF, onde um dos objetivos era a verificação do estado de sincronismo temporal, ou seja, as fases dos sinais originais e sinais reconstruídos. Com este objetivo, no cálculo da SNRSEG_NCCF os sinais originais e reconstruídos são divididos em segmentos. Para cada segmento do sinal reconstruído são então calculados: o valor máximo da correlação cruzada normalizada, NCCF, $R_{máx}$; a relação sinal ruído, SNR; o deslocamento do segmento, *d'*, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original; o deslocamento médio, $d'_{méd}$; e a própria relação sinal ruído segmental com auxílio da NCCF, ou seja, a medida SNRSEG-NCCF. Os deslocamentos, em número de amostras, indicam a defasagem entre os segmentos, onde *d'* indica a defasagem por segmento e $d'_{méd}$; indica a defasagem média para o sinal de acordo com o arranjo do experimento. Também foi calculado o valor médio das diferenças absolutas entre os deslocamentos para os segmentos sucessivos, $d'_{méd\_dif\_abs}$. Este parâmetro indica a defasagem média absoluta entre os segmentos consecutivos.

A avaliação da defasagem entre os sinais é realizada sobre os sinais fala do locutor adulto processados pelo sistema de análise - síntese WI utilizando os arquivos *casa2*.wav e *casa2*_sin_rec_pd.wav e sobre os sinais residuais *casa2*_res_pd.wav e *casa2*_res_rec_pd.wav que são utilizados respectivamente como sinais originais e sinais em teste. O processo de avaliação é feito considerando a inspeção visual sobre o gráfico das formas de onda

(amplitude *versus* tempo) do trecho sonoro nos sinais originais (referência) e nos sinais reconstruídos (em teste), e os resultados obtidos com o processamento dos mesmos sinais no algoritmo da SNRSEG-NCCF.

Assim, no processo de avaliação da defasagem são executados os seguintes itens:

(1) Avaliação da defasagem entre os sinais da fala (Seção 5.2.4.2.1 deste capítulo):
   (a) Resultados da aplicação do algoritmo SNRSEG-NCCF para *casa2*.wav e *casa2*_sin_rec_pd.wav;
   (b) Resultados da aplicação do algoritmo SNRSEG-NCCF para os trechos de sinais sonoros da fala extraídos dos arquivos *casa2*.wav e casa2_sin_rec_pd.wav;
   (c) Inspeção visual para os trechos dos sinais sonoros da fala extraídos dos arquivos *casa2*.wav e casa2_sin_rec_pd.wav;
   (d) Avaliação da defasagem entre os sinais da fala (*casa2*.wav e casa2_sin_rec_pd.wav).
(2) Avaliação da defasagem entre os sinais residuais da fala (Seção 5.2.4.2.2 deste capítulo):
   (a) Resultados da aplicação do algoritmo SNRSEG-NCCF para *casa2*_res_pd.wav e casa2_res_rec_pd.wav;
   (b) Resultados da aplicação do algoritmo SNRSEG-NCCF para os trechos dos sinais sonoros da fala extraídos dos arquivos *casa2*_res_pd.wav e casa2_res_rec_pd.wav;
   (c) Inspeção visual para os trechos de sinais sonoros da fala extraídos dos arquivos *casa2*_res_pd.wav e casa2_res_rec_pd.wav;
   (d) Avaliação da defasagem entre os sinais residuais.
(3) Avaliação geral da defasagem: sinais da fala e sinais residuais (*casa2*.wav e casa2_sin_rec_pd.wav, *casa2*_res_pd.wav e casa2_res_rec_pd wav) (Seção 5.2.4.2.3 deste capítulo).

No processamento do algoritmo para o cálculo da SNRSEG-NCCF os segmentos foram considerados com um tamanho fixo de 128 amostras ou 99 amostras, ou seja, M = 128 amostras ou M = 99 amostras. O valor de M = 99 amostras está relacionado com o valor do pitch médio para os sinais investigados nesta etapa. A intenção é verificar se o algoritmo SNRSEG-NCCF tem melhores resultados com o tamanho do segmento próximo do valor do pitch médio do sinal em avaliação. O valor M =128 amostras é considerado como um valor padrão na literatura para uma freqüência de amostragem de 8 kHz, neste trabalho a freqüência de amostragem de 11,025 kHz.

5.2.4.2.1 Avaliação da defasagem entre os sinais da fala para o sistema de análise – síntese WI (padrão)

(a) Resultados da aplicação do algoritmo SNRSEG-NCCF para *casa2*.wav e *casa2*_sin_rec_pd.wav

A seguir são apresentados os resultados da aplicação do algoritmo SNRSEG-NCCF para a expressão da fala "casa" locutor – adulto utilizando os arquivos *casa2*.wav e *casa2*_sin_rec_pd.wav. Na Tabela 5.4 são listados os resultados obtidos com o processamento do algoritmo da SNRSEG –NCCF utilizando o sinal da fala em *casa2*.wav como sinal original (referência) e o sinal reconstruído em *casa2*_sin_rec_pd.wav como sinal em teste. Em seguida são apresentados os gráficos esboçados a partir dos dados listados nessa Tabela.

Tabela 5.4 - Locutor – adulto - Sistema análise/síntese WI padrão – Mostra as medidas: deslocamento, correlação cruzada normalizada e a relação sinal ruído por segmento; o deslocamento médio e a relação sinal ruído segmental modificada com a NCCF entre os sinais da fala: ***casa2*.wav** (sinal original) e ***casa2*_sin_rec_pd.wav (**sinal em teste) para M = 128 amostras e M = 99 amostras.

Resultados: Programa SNRSEG-NCCF aplicado nos arquivos: sinal de referência: casa2.wav; e sinal em teste: casa2_sin_rec_pd.wav.						
	M =128 amostras			M = 99 amostras		
Índice do segmento	**Deslocamento** d' (número de amostras)	**Correlação cruzada Normalizada** R=>(NCCF)	**Relação sinal ruído (por segmento)** SNR (dB)	**Deslocamento** d' (número de amostras)	**Correlação cruzada Normalizada** R=>(NCCF)	**Relação sinal ruído (por segmento)** SNR (dB)
0	63	0,112983	0,032536	34	0,140938	0,046228
1	-51	0,21504	0,159791	-37	0,135875	0,046484
2	-20	0,634279	2,160136	21	0,228691	0,20646
3	-59	0,599175	1,793827	-20	0,705289	2,983333
4	13	0,649086	2,262883	-21	0,574451	1,736877
5	-57	0,677665	1,880417	13	0,564107	1,618236
6	-7	0,784183	4,026502	11	0,615356	1,214836
7	-10	0,945067	9,535499	-48	0,849724	4,388704
8	-7	0,888252	6,662916	-7	0,798272	4,224299
9	-5	0,978772	13,74118	-10	0,944353	9,493736
10	-5	0,97415	12,82067	-8	0,981487	14,13494
11	-4	0,98748	16,016109	-6	0,90715	7,313998
12	-5	0,985718	15,459266	-5	0,981658	14,39508
13	-4	0,944264	9,466081	-5	0,973389	12,6476
14	-4	0,990329	16,968496	-4	0,966752	11,76244
15	0	0,982508	14,597496	-5	0,977396	13,49237
16	1	0,950083	9,663156	-5	0,991821	17,76999
17	1	0,987015	15,861207	-4	0,950418	10,04731
18	1	0,979625	13,679173	-4	0,994476	19,52893
19	0	0,975461	13,094779	0	0,928072	8,381985
20	0	0,984769	15,117928	0	0,972231	12,61482
21	0	0,987384	15,878607	1	0,986244	15,47692
22	-1	0,966624	11,712086	1	0,986467	15,67596
23	0	0,982493	14,135993	0	0,982051	14,31692
24	-1	0,982643	14,400113	1	0,996058	20,54045
25	0	0,990849	13,896808	0	0,99014	16,96914
26	0	0,989787	11,622254	0	0,983598	14,77695
27	0	0,986431	14,130206	0	0,991184	17,55452
28	0	0,96232	9,994581	0	0,982825	14,37315

29	2	0,971599	12,249097	-2	0,986015	15,29826
30	1	0,965188	10,722023	0	0,985902	15,04602
31	2	0,98617	15,47997	-1	0,986259	15,25402
32	4	0,977682	13,471747	-1	0,986696	14,14215
33	3	0,961895	10,6307	0	0,988119	11,3458
34	4	0,963651	11,02941	0	0,984937	11,32479
35	1	0,95991	11,010293	0	0,988675	14,66065
36	2	0,978576	13,638454	1	0,977332	13,44457
37	1	0,979818	12,076218	1	0,972673	10,30873
38	1	0,988957	14,585525	0	0,971082	11,63872
39	1	0,961858	10,986402	1	0,967761	11,91232
40	2	0,981024	14,140146	2	0,985376	14,27601
41	2	0,981842	14,421568	3	0,984153	15,0112
42	2	0,990671	16,831243	4	0,973021	12,49269
43	1	0,987403	15,931473	3	0,961245	11,08201
44	1	0,985936	15,526363	2	0,952626	10,16779
45	-1	0,989386	16,530733	1	0,973892	12,63203
46	0	0,975177	12,580657	1	0,985771	14,09582
47	0	0,982806	14,405985	2	0,983644	14,62859
48	-1	0,97716	12,249892	1	0,978249	11,47599
49	-2	0,984778	15,192934	1	0,992671	14,82987
50	-2	0,985359	13,636078	1	0,98041	11,26957
51	-2	0,986574	15,733445	2	0,979979	14,01818
52	-3	0,970774	12,034255	2	0,983115	14,39034
53	-2	0,981743	12,232312	2	0,981929	14,44059
54	-3	0,98538	10,784039	2	0,99168	17,21989
55	7	0,764287	3,391651	2	0,987769	16,0367
56	-25	0,904927	7,217159	1	0,993986	18,9619
57	5	0,601646	0,461142	1	0,986325	15,62921
58	-	-	-	-1	0,989772	16,91413
59	-	-	-	-1	0,984721	14,55442
60	-	-	-	0	0,992068	16,60822
61	-	-	-	0	0,982732	14,37529
62	-	-	-	0	0,988065	14,76772
63	-	-	-	-1	0,987863	15,10654
64	-	-	-	-2	0,987451	15,95455
65	-	-	-	-2	0,986359	13,56892
66	-	-	-	-2	0,990376	16,61703
67	-	-	-	-4	0,976094	13,19224
68	-	-	-	-2	0,969277	12,17317
69	-	-	-	-3	0,980732	9,844407
70	-	-	-	-2	0,994849	14,22989
71	-	-	-	2	0,916327	6,256062
72	-	-	-	-32	0,439651	0,911501
73	-	-	-	-25	0,883368	6,438326
74	-	-	-	37	0,630322	0,688937
d'_med = -2 amostras		**d'_méd_dif_abs= 9 amostras**		**d'_med = -1 amostras**		**d'_méd_dif_abs= 6 amostras**
SNRSEG-NCCF = 11.378476 dB				**SNRSEG-NCCF = 11.959578 dB**		
Segmentos nulos = 0		Segmentos abaixo do limiar = 0		Segmentos nulos = 0		Segmentos abaixo do limiar = 0

A partir dos dados da Tabela 5.4 foram esboçados os gráficos mostrados nas Figuras 5.4, 5.5, 5.6, 5.7, 5.8, 5.9 e 5.10. As Figuras 5.4 e 5.5 mostram os deslocamentos, d' (em número de amostras ou em amostras), que foram obtidos para cada segmento do arquivo em teste com a

aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 128 e M = 99 amostras respectivamente.

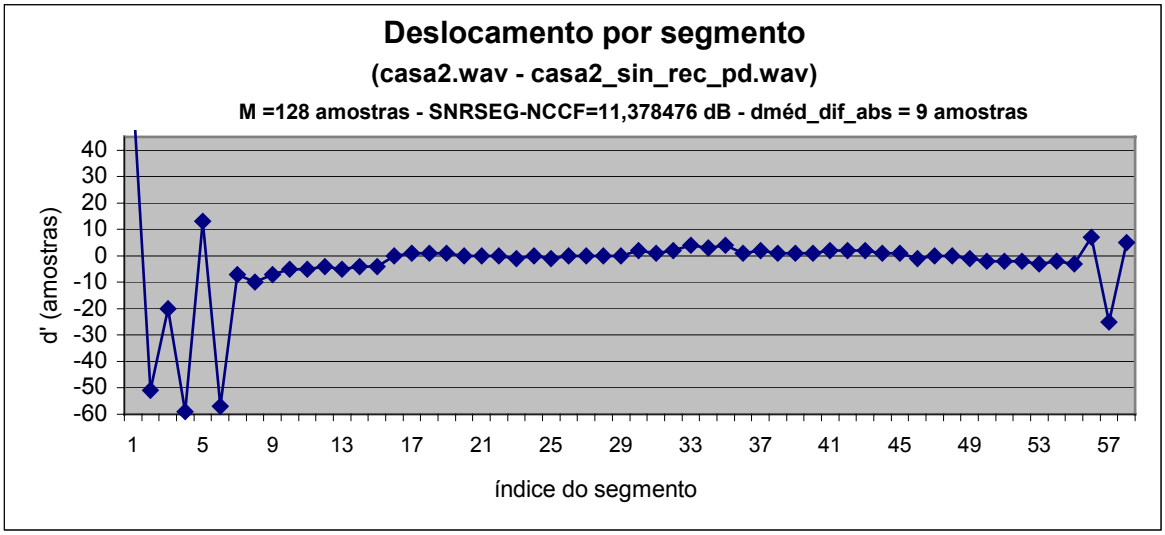

Figura 5.4 – Deslocamentos em número de amostras, d', para os segmentos do sinal reconstruído, *casa2*_sin_rec_pd.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original, *casa2*.wav. Resultados obtidos para M =128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

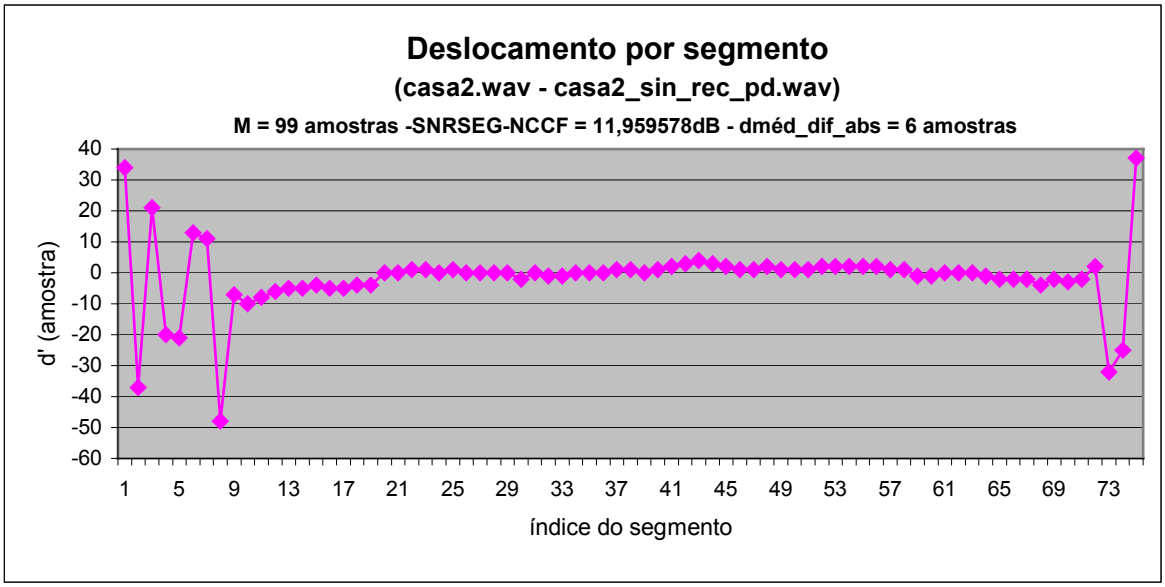

Figura 5.5 – Deslocamentos em número de amostras, d', para os segmentos do sinal reconstruído, *casa2*_sin_rec_pd.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original, *casa2*.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

As Figuras 5.4 e 5.5 mostram que os deslocamentos, d', são menores nas regiões onde os sinais são sonoros e tem maior potência. Nas regiões mais ruidosas os deslocamentos são maiores. Quanto menor for a diferença entre os deslocamentos para segmentos consecutivos menor será a diferença de fase entre os sinais. Observa-se que existem regiões onde os deslocamentos são constantes e diferentes de zero e regiões onde os deslocamentos são constantes e próximo de zero. Pode-se dizer que os sinais estão quase em fase em cada uma dessas regiões. Considerando o sinal completo pode-se dizer que nos trechos inicial e final a fase é variável e que ao longo do trecho central, na maior parte do sinal, a fase é quase constante.

As Figuras 5.6 e 5.7 mostram as correlações cruzadas normalizadas, R (NCCF), entre os segmentos no arquivo *casa2*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd.wav calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para segmentos com M = 128 e M = 99 amostras, respectivamente.

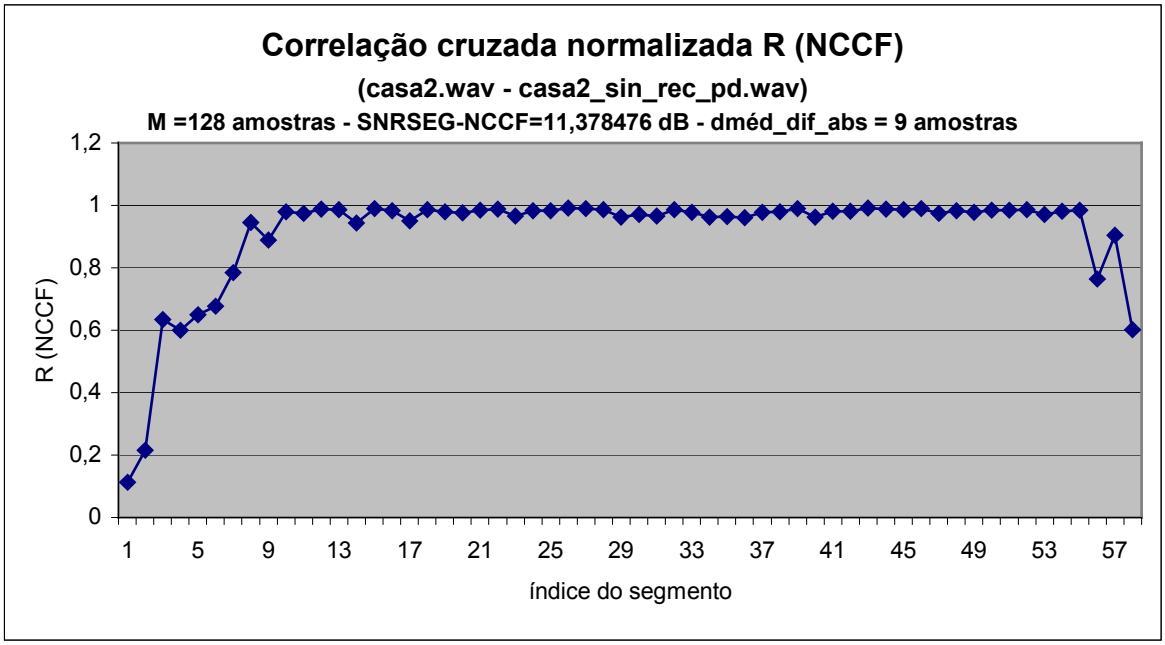

Figura 5.6 – Correlação cruzada normalizada, R (NCCF), entre os segmentos no arquivo *casa2*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd.wav. Resultados obtidos para M = 128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

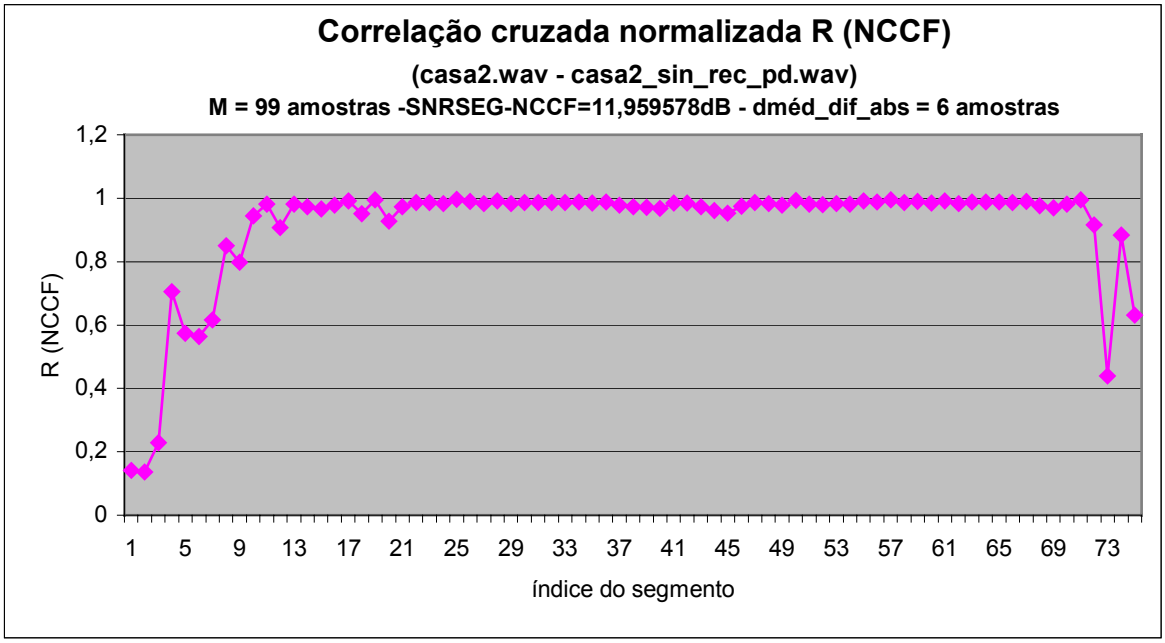

Figura 5.7 – Correlação cruzada normalizada, R (NCCF), entre os segmentos no arquivo *casa2*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

As Figuras 5.6 e 5.7 mostram que as correlações cruzadas normalizadas são maiores (próximas do valor 1,0) nas regiões onde o sinal é mais potente e o pitch bem definido , ou seja, nas regiões sonoras. Neste caso a variação de R é pequena entre segmentos consecutivos.

Nos extremos onde o sinal é mais ruidoso e com uma potência menor as correlações cruzadas normalizadas são menores e com grande variação entre os segmentos consecutivos.

Pode existir também para o primeiro e último segmento o efeito de borda devido ao algoritmo da SNRSEG-NCCF, quando a janela deslizante enquadra M+M/2 ou M/2+M amostras e as outras M/2 posições da janela são completadas com valores iguais a zero. Assim os valores das correlações cruzadas normalizadas devem ser desconsiderados.

As Figuras 5.8 e 5.9 mostram as relações sinal ruído, SNR (dB), entre os segmento no arquivo *casa2*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 128 e M = 99 amostras, respectivamente.

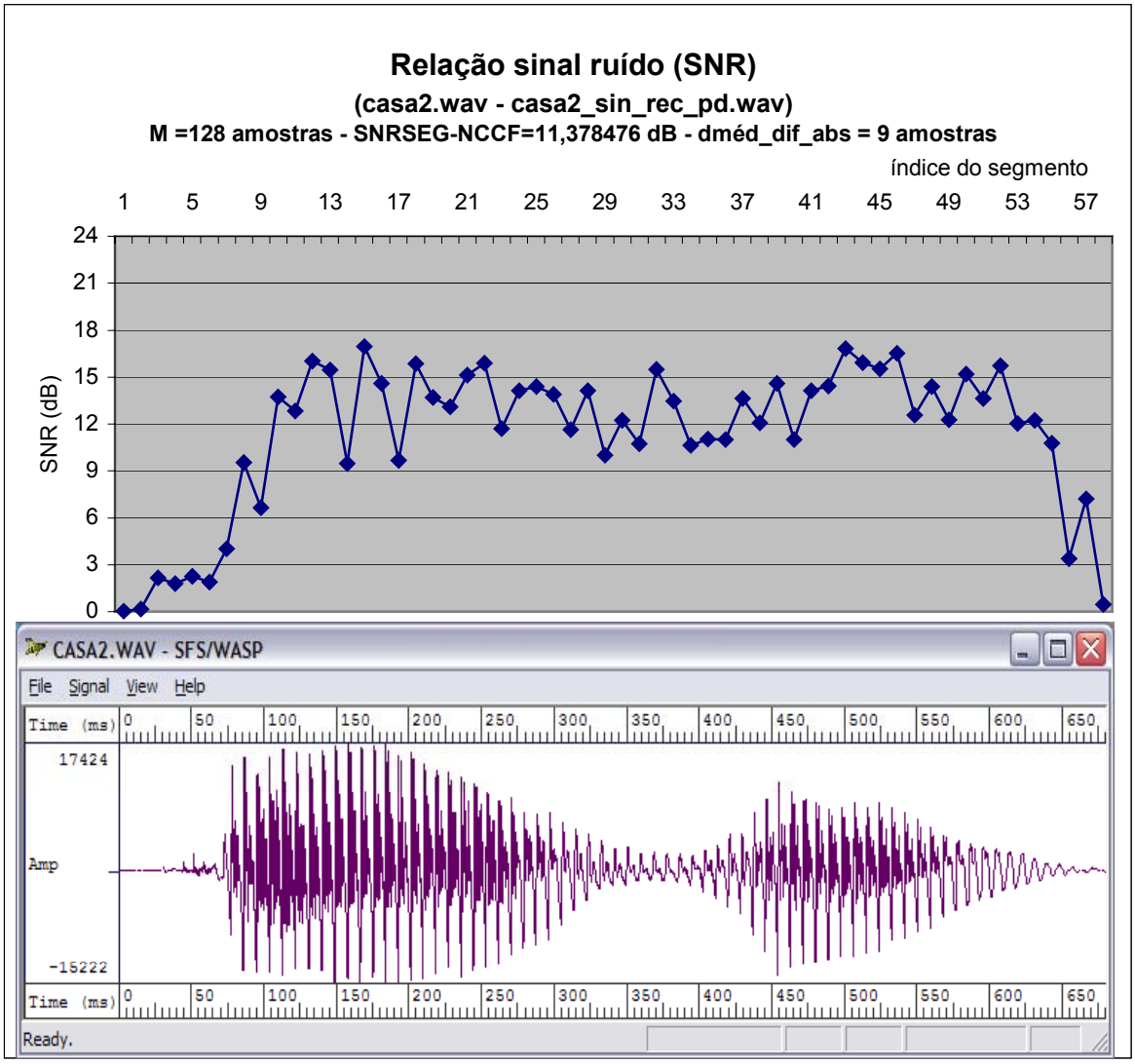

Figura 5.8 – Relação sinal ruído, SNR (dB), entre os segmento no arquivo *casa2*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd.wav. Para a visualização da posição relativa dos segmentos é mostrado também o gráfico das formas de onda amplitude x tempo para o sinal *casa2*.wav. Estes resultados foram obtidos para M = 128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

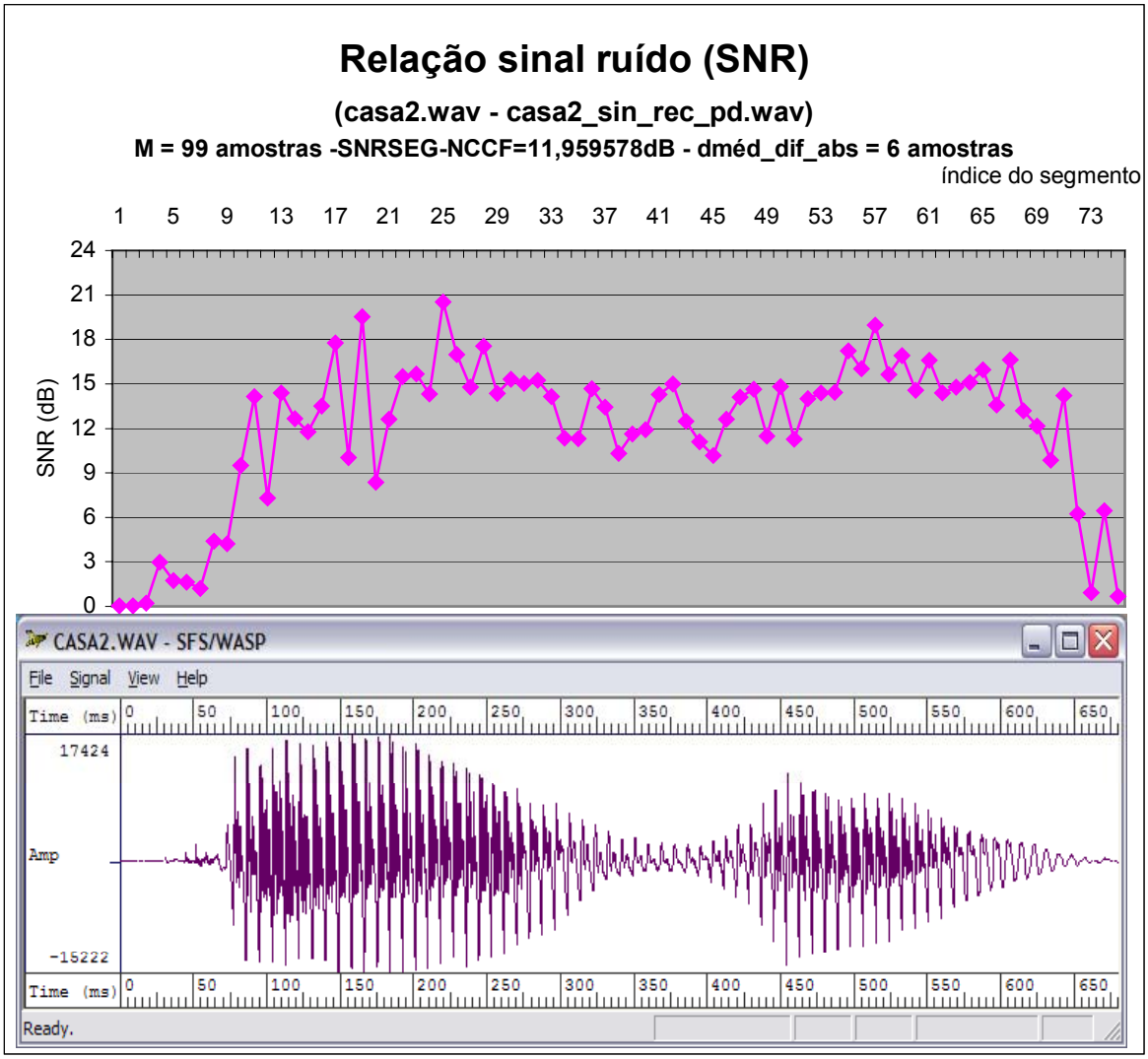

Figura 5.9 – Relação sinal ruído, SNR (dB), entre os segmento no arquivo *casa2*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd.wav. Para a visualização da posição relativa dos segmentos é mostrado também o gráfico das formas de onda amplitude x tempo para o sinal *casa2*.wav. Estes resultados foram obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

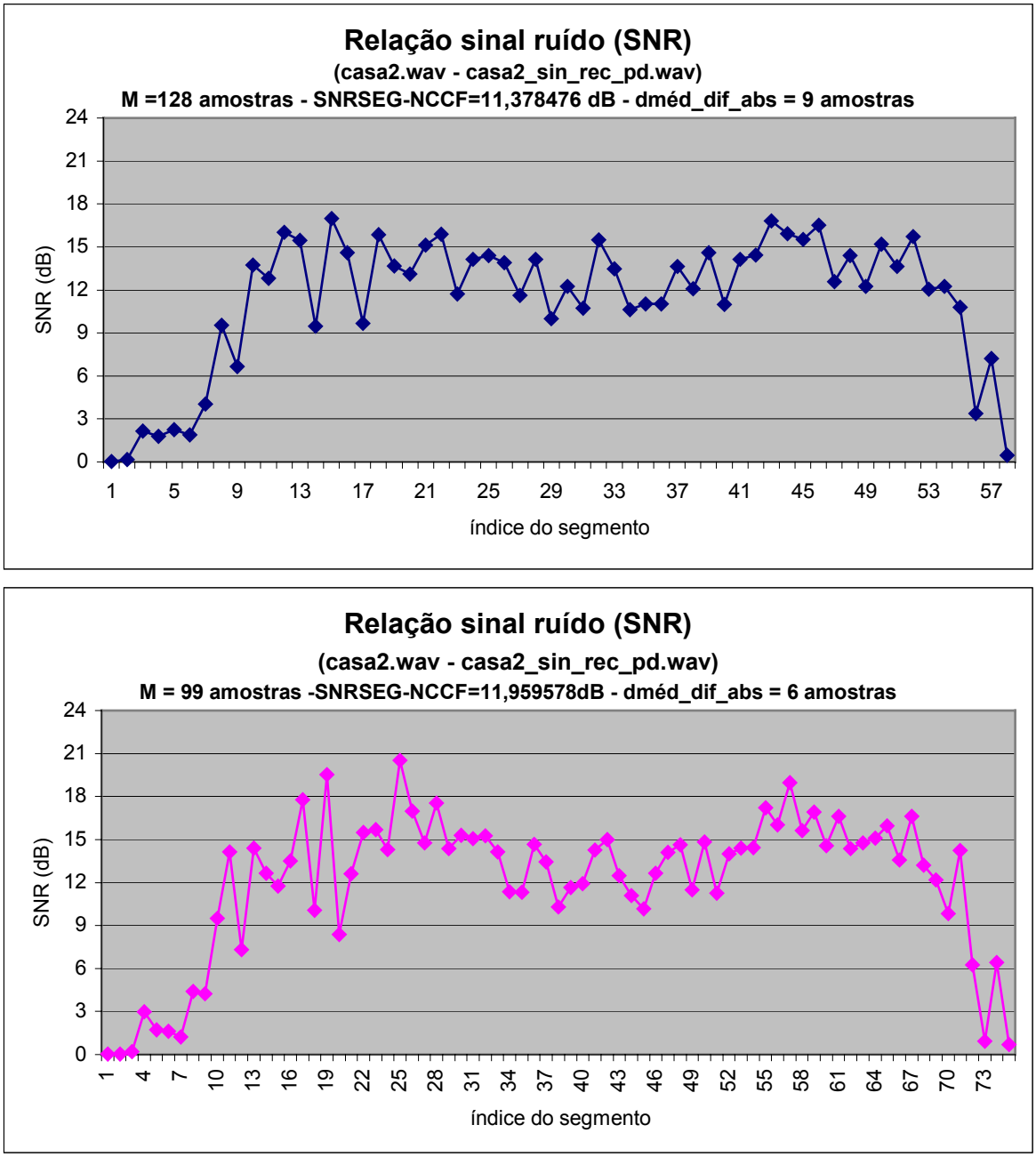

Figura 5.10 - Repetição das Figuras 5.8.e 5.9 para a comparação entre o gráfico das SNR's entre os segmentos para M = 128 no gráfico superior e M = 99 amostras no gráfico inferior.

As Figuras 5.8, 5.9 e 5.10 mostram que a SNR por segmento é maior nas regiões onde o sinal é sonoro e mais potente. Isto mostra que nestas regiões os sinais são mais similares, o que é comprovado pelos valores da correlação cruzada normalizada nas Figuras 5.6 e 5.7.

Observa-se também que na Figura 5.9 (para M = 99 amostras) a SNR tem valores maiores para alguns segmentos, o que resulta em média em uma medida SNRSEG-NCCF maior do que no caso para M = 128 amostras. Isto é conseqüência do melhor enquadramento das

formas de onda cíclicas pelos segmentos, ou seja, os comprimentos das formas de onda que se repetem no tempo são próximos do tamanho M do segmento.

Nos extremos onde o sinal é mais ruidoso e com uma potência menor, as SNR's são menores.

Aqui também pode ocorrer, para o primeiro e último segmento, o efeito de borda devido ao algoritmo da SNRSEG-NCCF, quando a janela deslizante enquadra M+M/2 ou M/2+M amostras e as outras M/2 posições da janela são completadas com valores iguais a zero. Assim os valores das SNR's nos dois extremos devem ser desconsiderados.

(b) Resultados da aplicação do algoritmo SNRSEG-NCCF para os trechos de sinais sonoros da fala extraídos dos arquivos *casa2*.wav e casa2_sin_rec_pd.wav

A seguir são apresentados os resultados da aplicação do algoritmo SNRSEG-NCCF para a expressão da fala "casa" locutor – adulto utilizando os trechos com segmentos sonoros da fala extraídos dos arquivos *casa2*.wav (sinal original - referência) e casa2_sin_rec_pd.wav (sinal reconstruído - em teste). Os resultados são apresentados nos gráficos nas Figuras 5.12, 5.13, 5.14, 5.15, 5.16 e 5.17.

A Figura 5.11 mostra a localização dos segmentos sonoro da fala considerados nos arquivo casa2.wav e casa2_sin_rec_pd.wav. Os trechos com os sinais sonoros foram extraídos em posições correspondentes nos dois arquivos.

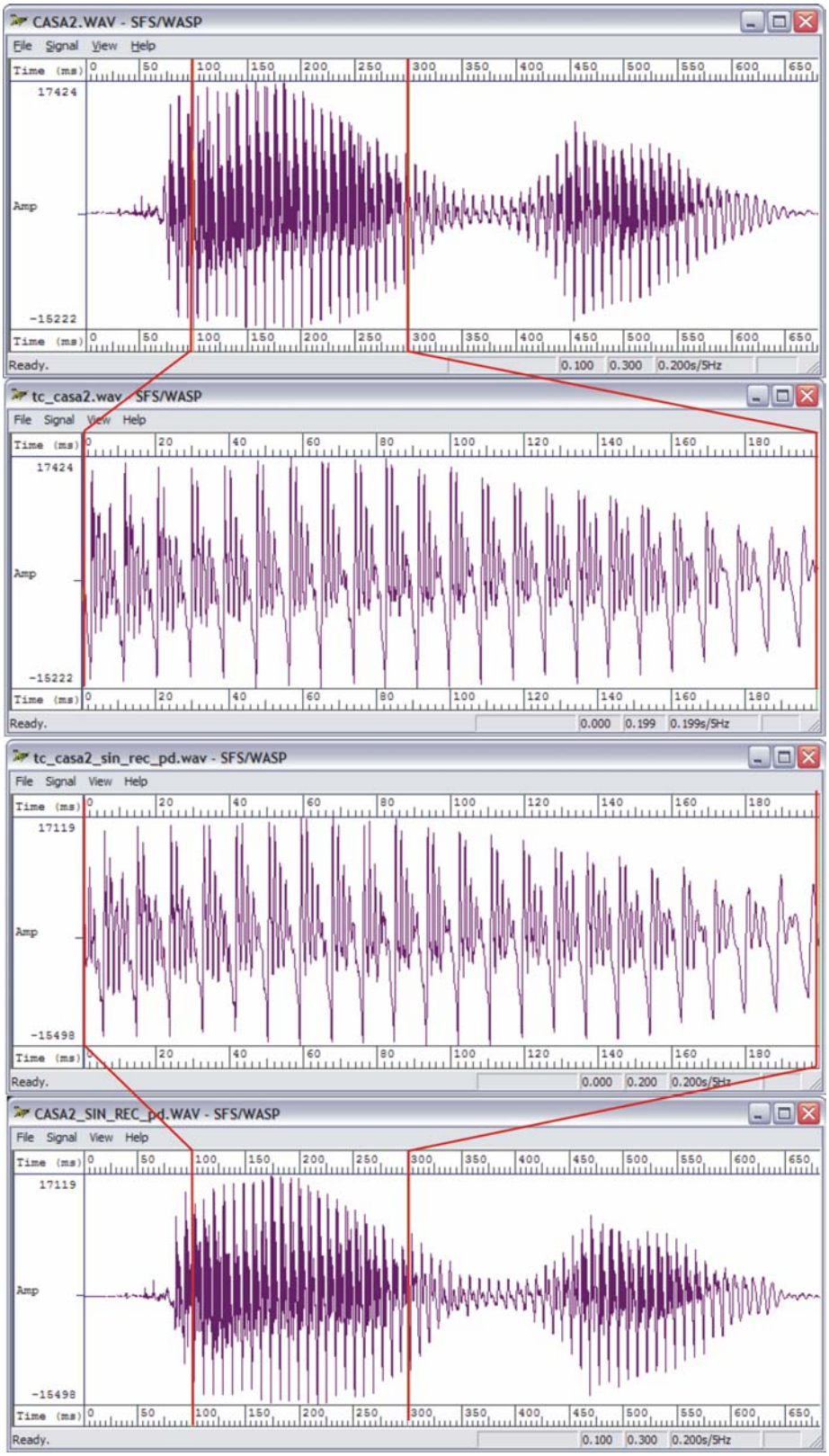

Figura 5.11- A parte superior da figura mostra o sinal da fala original no arquivo casa2.wav. Na parte inferior mostra o sinal da fala reconstruído no arquivo casa2_sin_rec_pd.wav. A parte intermediária mostra as ampliações dos trechos sonoros correspondentes, tc_casa2.wav e tc_casa2_sin_rec.wav, extraídos, respectivamente, do sinal da fala original e do sinal reconstruído.

As Figuras 5.12 e 5.13 mostram os deslocamentos, d' (em amostras), que foram obtidos para cada segmento do arquivo em teste (trecho sonoro em casa2_sin_rec_pd.wav) em relação ao arquivo original (trecho sonoro em casa2.wav) com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 128 e M = 99 amostras respectivamente.

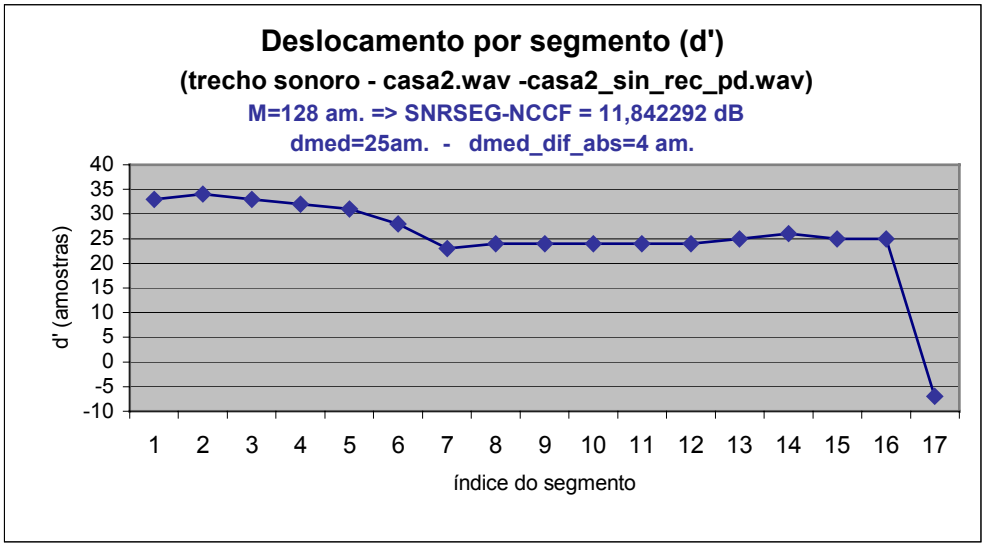

Figura 5.12 – Deslocamentos em amostras, d', para os segmentos do trecho sonoro extraído do sinal reconstruído, no arquivo *casa2*_sin_rec_pd.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no trecho sonoro correspondente extraído do sinal original *casa2*.wav para M = 128 amostras.

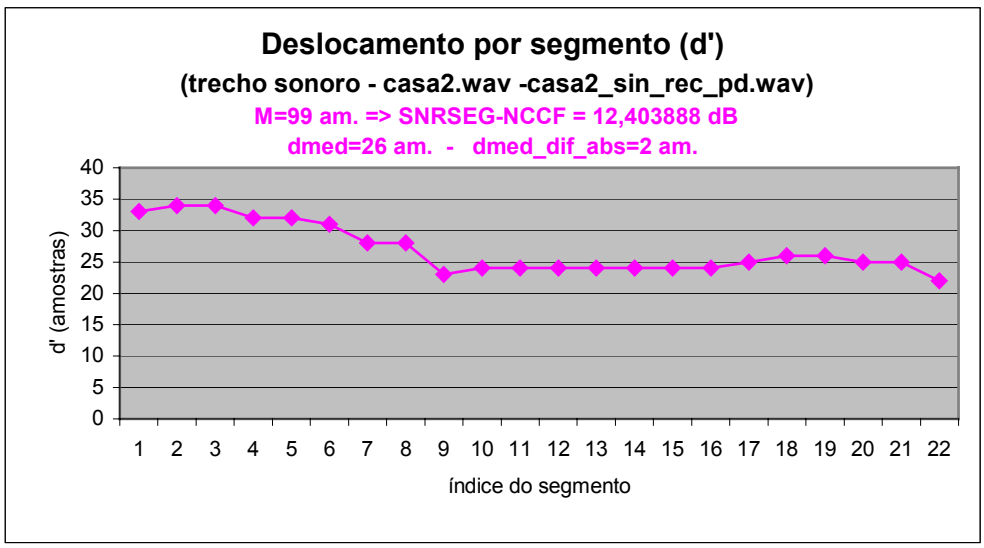

Figura 5.13 – Deslocamentos em amostras, d', para os segmentos do trecho sonoro extraído do sinal reconstruído, no arquivo *casa2*_sin_rec_pd.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no trecho sonoro correspondente extraído do sinal original, no arquivo *casa2*.wav para M = 99 amostras.

As Figuras 5.12 e 5.13 mostram que os deslocamentos, d', são menores nas regiões onde os sinais são sonoros e com maiores potências. Nas regiões mais ruidosas os deslocamentos são maiores. Quanto menor for a diferença entre os deslocamentos para os segmentos consecutivos menor será a diferença de fase entre os sinais. Observa-se que existem regiões onde os deslocamentos são constantes e diferentes de zero e regiões onde os deslocamentos são constantes e próximo de zero. Pode-se dizer que os sinais estão quase em fase em cada uma dessas regiões. Considerando o sinal completo pode-se dizer que na região inicial a fase é variável e que ao longo das regiões central e final, na maior parte do sinal, a fase é quase constante.

As Figuras 5.14 e 5.15 mostram as correlações cruzadas normalizadas, R (NCCF), entre os segmentos do trecho sonoro extraído no arquivo *casa2*.wav, e os segmentos mais similares próximos às posições correspondentes no trecho sonoro extraído no arquivo *casa2*_sin_rec_pd.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 128 e M =99 amostras respectivamente.

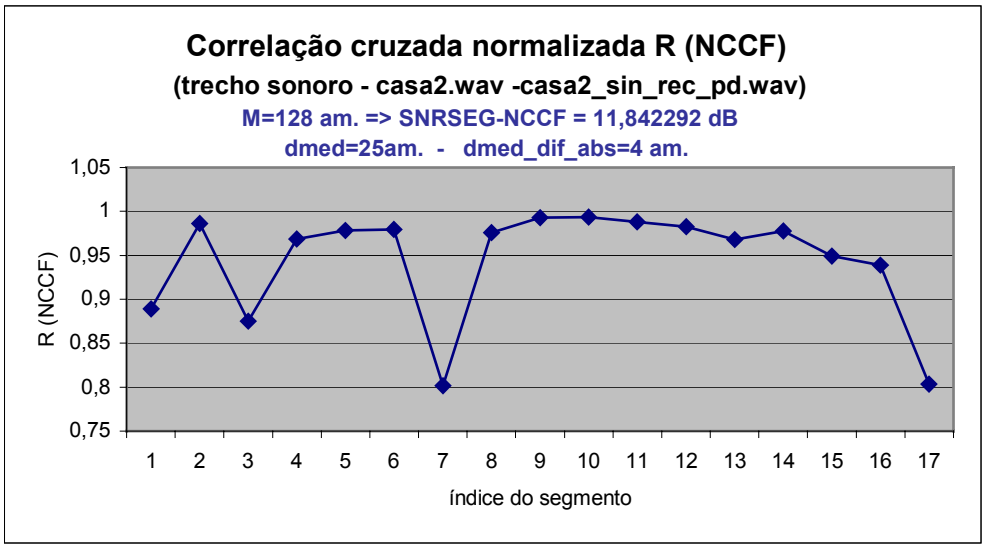

Figura 5.14 – Correlação cruzada, R (NCCF), entre os segmentos do trecho sonoro extraído no arquivo *casa2*.wav, e os segmentos mais similares próximos às posições correspondentes do trecho sonoro extraído no arquivo *casa2*_sin_rec_pd.wav. Resultados obtidos para M = 128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

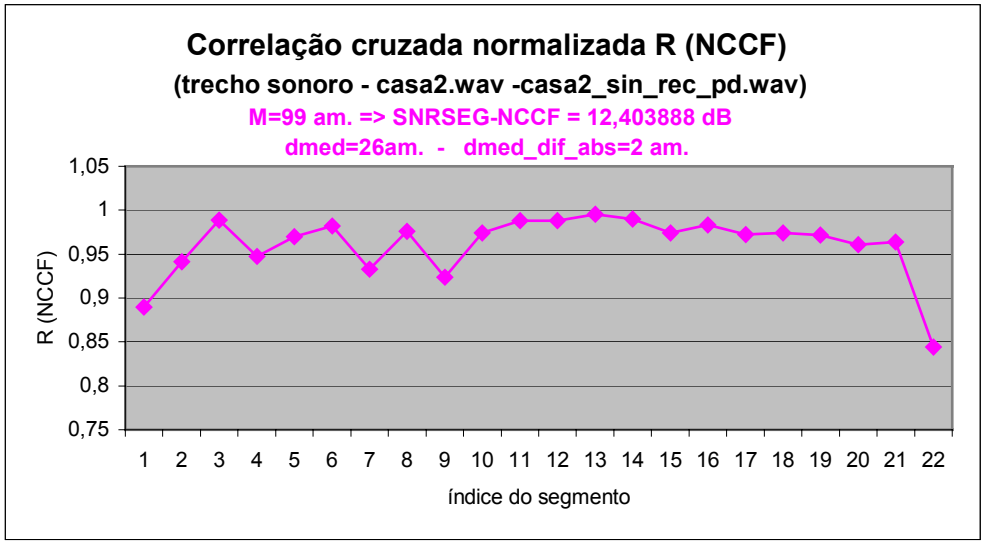

Figura 5.15 – Correlação cruzada, R (NCCF), entre os segmentos do trecho sonoro extraído no arquivo *casa2*.wav, e os segmentos mais similares próximos às posições correspondentes do trecho sonoro extraído no arquivo *casa2*_sin_rec_pd.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

As Figuras 5.14 e 5.15 mostram que as correlações cruzadas normalizadas são mais próximas do valor 1,0 para o caso onde M = 99 amostras. Neste caso a variação de R é menor entre segmentos consecutivos comparando com o caso quando M = 128 amostras.
Pode existir também para o primeiro e último segmento o efeito de borda devido ao algoritmo da SNRSEG-NCCF, quando a janela deslizante enquadra M+M/2 ou M/2+M amostras e as outras M/2 posições da janela são completadas com valores iguais a zero. Assim os valores das correlações cruzadas normalizadas nos segmentos extremos podem ser desconsiderados.

As Figuras 5.16 e 5.17 mostram as relações sinal ruído, SNR (dB), entre os segmentos do trecho sonoro extraído no arquivo *casa2*.wav e os segmentos correspondentes com a maior similaridade no trecho sonoro extraído no arquivo *casa2*_sin_rec_pd.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 128 e M = 99 amostras respectivamente.

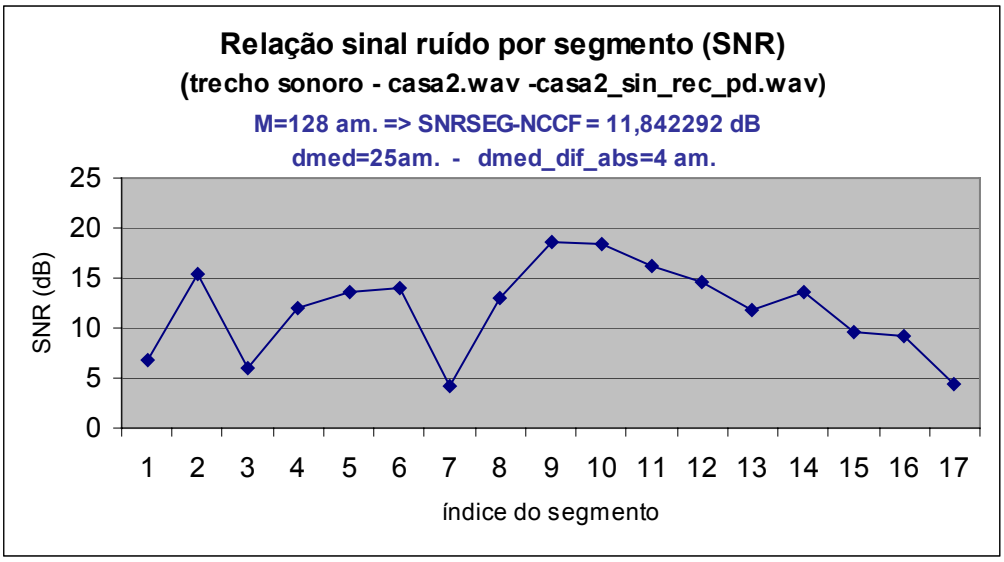

Figura 5.16 – Relação sinal ruído, SNR (dB), entre os segmentos do trecho sonoro extraído no sinal *casa2*.wav e os segmentos correspondentes com a maior similaridade no trecho sonoro extraído no sinal *casa2*_sin_rec_pd.wav. Resultados obtidos para M = 128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

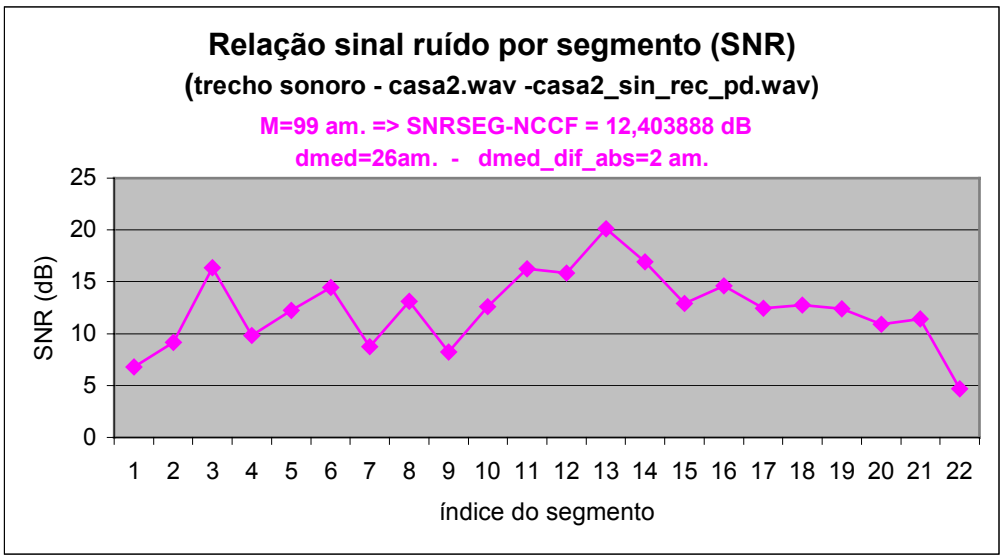

Figura 5.17 – Relação sinal ruído, SNR (dB), entre os segmentos do trecho sonoro extraído no sinal *casa2*.wav e os segmentos correspondentes com a maior similaridade no trecho sonoro extraído no sinal *casa2*_sin_rec_pd.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

Na Figura 5.17, para o caso de M = 99 amostras, a seqüência de valores da SNR por segmento tem uma variação menor do que na Figura 5.16, para o caso onde M =128 amostras. Isto mostra que os segmentos com M = 99 amostras são mais similares, o que é comprovado pelos valores da correlação cruzada normalizada nas Figuras 5.14 e 5.15.

Observa-se também que na Figura 5.16 (para M = 99 amostras) a SNR tem valores maiores para alguns segmentos, o que resulta em média em uma medida SNRSEG-NCCF maior do que no caso para M = 128 amostras. Isto é conseqüência do melhor enquadramento das formas de onda cíclicas pelos segmentos, ou seja, os comprimentos das formas de onda que se repetem no tempo são próximos do tamanho M do segmento.

Aqui também pode ocorrer, para o primeiro e último segmento, o efeito de borda devido ao algoritmo da SNRSEG-NCCF, quando a janela deslizante enquadra M+M/2 ou M/2+M amostras e as outras M/2 posições da janela são completadas com valores iguais a zero. Assim os valores das SNR's nos dois extremos podem ser desconsiderados.

(c) Inspeção visual para os trechos de sinais sonoros da fala extraídos dos arquivos *casa2*.wav *e casa2*_sin_rec_pd.wav

A Figura 5.11 mostra os trechos sonoros segmentados dos sinais nos arquivos *casa2*.wav e *casa2*_sin_rec_pd.wav que foram utilizados para efetuar a análise da defasagem entre dos sinais.

Através do software SFS/WASP[5.1] foram levantados por inspeção direta sobre as formas de ondas dos sinais, o valor do pitch (ciclo a ciclo) e a diferença temporal (a defasagem) entre os picos dos pulsos dos ciclos de pitch, que foram convertidas para número de amostras, considerando a taxa de amostragem de 11025 Hz para os sinais. Os resultados são mostrados na Tabela 5.5.

[5.1] SFS/WASP version 1.06 (3/10/2000) (c) 2000 Mark Huck vale University College London. http://www.phon.ucl.ac.uk/resource/sfs/

Tabela 5.5 – Valores obtidos a partir da inspeção visual para os trechos sonoros segmentados dos arquivos casa2.wav e casa2_sin_rec_pd.wav mostrados na Figura 5.11: pitch (Hz) e defasagem (amostras) para freqüência de amostragem de 11.025 Hz.

Parâmetros obtidos com a inspeção visual			
Índice dos ciclos do pitch	Trecho sonoro do sinal original casa2.wav	Trecho sonoro do sinal reconstruído casa2_sin_rec_pd.wav	Defasagem entre os picos dos pulsos dos ciclos do pitch (amostras)
	Pitch (Hz)	Pitch (Hz)	
1	111	111	33
2	111	111	33
3	108	111	33
4	114	114	33
5	111	114	33
6	111	117	33
7	117	117	22
8	114	117	33
9	120	117	22
10	114	114	22
11	117	117	33
12	114	114	22
13	117	114	33
14	114	117	33
15	117	114	22
16	117	117	22
17	114	111	33
18	117	114	33
19	114	114	22
20	117	111	33
21	108	114	33
22	117	114	22
23	-	-	22
--	$Pitch_{méd}$ = 114,27 Hz	$Pitch_{méd}$ = 114,27 Hz	$Defasagem_{média}$ = 29 amostras
--	$Pitch_{méd}$= 96,48 amostras	$Pitch_{méd}$ = 96,48 amostras	

A Figura 5.18 mostra o gráfico da defasagem entre os picos de pulsos do pitch obtidos por inspeção visual de acordo com a Tabela 5.5 para a comparação com o gráfico da Figura 5.13

onde é mostrado o gráfico do deslocamento por segmento para os mesmos trechos do sinal da fala, nos arquivos *casa2*.wav e *casa2*_sin_rec_pd.wav, que foram obtidos com a aplicação do algoritmo da SNRSEG-NCCF para M = 99 amostras.

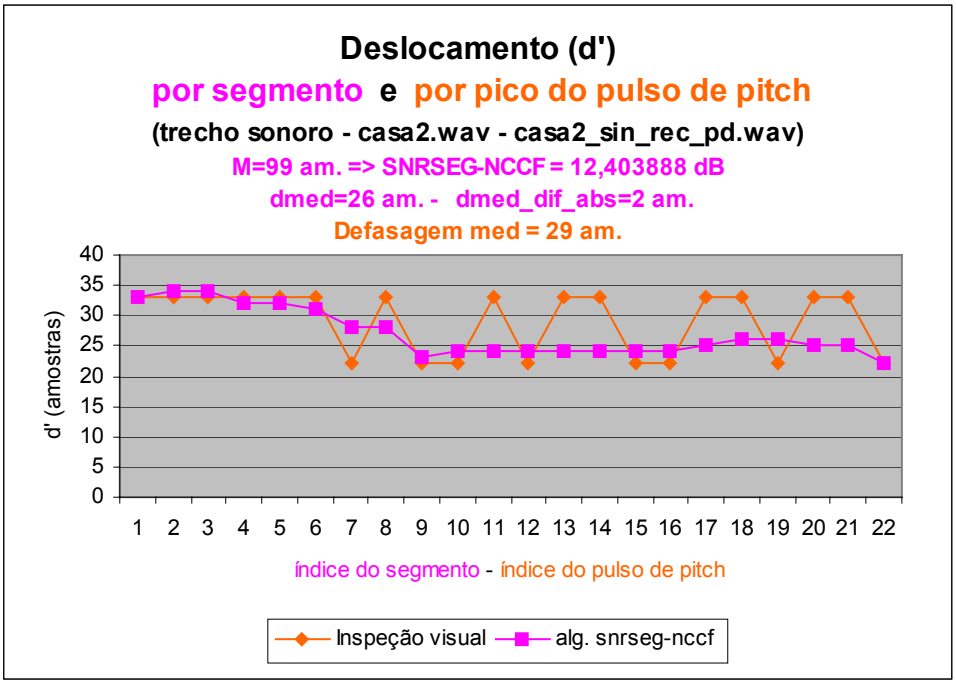

Figura 5.18.- Gráfico para a comparação entre: ***(a)*** *o* ***deslocamento por segmento*** *para o trecho do sinal da fala com a aplicação do algoritmo da SNRSEG-NCCF (M = 99 amostras)* (na cor magenta)*; e* ***(b)*** *a* ***defasagem entre os picos dos pulsos de pitch*** *obtidos por inspeção visual no mesmo trecho do sinal da fala de acordo com a Tabela 5.5* (na cor laranja). Estes resultados foram obtidos a partir do trecho extraído no sinal da fala correspondente ao trecho sonoro nos sinais da fala reconstruídos do arquivo *casa2*_sin_rec_pd.wav e nos sinais originais no arquivo *casa2*.wav.

A Figura 5.18 mostra que, para o mesmo trecho sonoro da fala, os valores para o **deslocamento por segmento** (obtidos *com a aplicação do algoritmo da SNRSEG-NCCF – M = 99 amostras* - na cor magenta) e os valores para a ***defasagem entre os picos dos pulsos de pitch*** *(obtidos por inspeção visual* - na cor laranja) estão dentro de uma faixa semelhante, onde o deslocamento médio foi de 26 amostras e a defasagem media foi de 29 amostras.

(d) Avaliação da defasagem entre os sinais da fala (*casa2*.wav *e casa2*_sin_rec_pd.wav)

Para avaliação da defasagem entre os sinais da fala, nos arquivos casa2.wav e casa2_sin_rec_pd.wav, são considerados os resultados obtidos nos itens (a), (b) e (c) que abordam a aplicação do algoritmo SNRSEG-NCCF e a inspeção visual.

Os dados da inspeção visual para trecho sonoro da fala, mostrados na Tabela 5.5 e na Figura 5.18 (no item (c)), indicam que:

- O pitch manteve-se variando na faixa de 108 a 120 Hz para o sinal de referência e de 111 a 117 Hz para o sinal em teste;
- Comparando o valor do pitch para as formas de ondas dos ciclos de pitch correspondentes nos dois sinais observa-se coincidência em quase 50% e o restante diferem entre 3 e 6 Hz o que corresponde a uma variação aproximada em torno do $Pitch_{méd}$ ($Pitch_{méd}$ = 96 amostras) de até 6 amostras aproximadamente para a freqüência de amostragem f_s = 11.025 Hz;
- A diferença temporal (ou defasagem em amostras) entre os picos dos pulsos do pitch ficou na média de 29 amostras, variando entre 22 e 33 amostras, sendo que grande parte (60,87 %) permaneceu em torno de 33 amostras e a outra parte (39,13 %) em torno de 22 amostras. Observa-se uma seqüência de segmentos com defasagem constante em 33 amostras e outra seqüência de segmentos com defasagem variável entre 22 e 33 amostras.

Os dados obtidos com a aplicação do algoritmo SNRSEG-NCCF são mostrados nas Figuras 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.12, 5.13, 5.14, 5.15, 5.16 e 5.17. *Estes dados indicam que*:

Para o sinal da fala (arquivo completo) – (no item (a)):

- O deslocamento $d_{méd\_dif\_abs}$ foi de 6 amostras para M = 99 amostras e de 9 amostras para M = 128 amostras, indicando que em média os segmentos foram deslocados consecutivamente de 6 amostras e 9 amostras para o cálculo da SNRSEG-NCCF com M = 99 amostras e M = 128 amostras respectivamente;
- A medida SNRSEG-NCCF entre os sinais da fala (referência e reconstruído) foi de 11,959578 dB (para M = 99 amostras) e de 11,378476 dB (para M = 128 amostras);
- A maioria dos resultados, conforme as Tabelas 5.1, 5.2 e 5.3 na Seção 5.2.4.1 deste capítulo, também mostra maiores valores para a medida SNRSEG-NCCF e menores valores para $d_{méd\_dif\_abs}$ para M = 99 amostras.

Assim p*ara o sinal da fala (arquivo completo)* pode-se dizer que:

- Os sinais apresentaram regiões onde os deslocamentos para os segmentos são variáveis, quase constantes e constantes: *deslocamentos variáveis* – trechos sem sincronismo, para sinais sonoros com baixa potência; *deslocamentos quase constantes* - pouca variação entre os deslocamentos sucessivos, indicando sinais com defasagem quase constante; e *deslocamentos constantes*: pequenos trechos com sinais com defasagem constante.
- De forma geral os sinais não estão em fase, ou seja, têm fase variável. Para segmentos com M = 99 amostras o algoritmo apresentou os melhores resultados.

Para o trecho sonoro do sinal da fala (no item (b)):

- O deslocamento $d_{méd\_dif\_abs}$ foi de 2 amostras para M = 99 amostras e de 4 amostras para M = 128 amostras, indicando que em média os segmentos foram deslocados consecutivamente de

2 amostras e 4 amostras para o cálculo da SNRSEG-NCCF com M = 99 amostras e M = 128 amostras respectivamente;
- O deslocamento $d_{méd}$ foi de 26 amostras para M = 99 amostras e de 25 amostras para M = 128 amostras, indicando que em média os segmentos foram deslocados de 26 amostras e 25 amostras para o cálculo da SNRSEG-NCCF com M = 99 amostras e M = 128 amostras respectivamente;
- A medida SNRSEG-NCCF entre os trechos sonoros dos sinais da fala (referência e reconstruído) foi de 12,403888 dB (para M = 99 amostras) e de 11,842292 dB (para M = 128 amostras).

Assim p*ara o trecho sonoro do sinal da fala* pode-se dizer que:
- Os trechos de sinais sonoros confirmam os melhores resultados para segmentos com M = 99 amostras. Comparando os resultados do algoritmo SNRSEG-NCCF com os resultados da inspeção visual percebe-se que novamente $d_{méd}$ = 26 amostras (M = 99 amostras) também está mais próximo do defasamento médio = 29 amostras.

***Considerações finais (seção* 5.2.4.2.1)** - *Com os dados apresentados para os sinais da fala,* nos arquivos *casa2.wav e casa2_sin_rec_pd.wav pode-se dizer que:*
- De forma geral os sinais da fala não estão em sincronismo, ou em fase;
- Nos trechos sonoros com maiores potências o algoritmo mostra que os sinais apresentam regiões com defasagem quase constante e regiões com defasagem constante entre os segmentos. E nos trechos sonoros com baixas potências, ou sinais mais ruidosos, o algoritmo mostra que não existe sincronismo (defasagem variável);
- O algoritmo apresentou melhores resultados (medida SNRSEG-NCCF, $d_{méd}$, e $d_{méd\_dif\_abs}$) entre os trechos sonoros utilizando M = 99 amostras próximo ao valor do pitch médio de 96 amostras. Isto é um indicativo, a ser investigado nas pesquisas futuras, de que se houver sincronismo entre o tamanho M dos segmentos com o valor do pitch médio, o algoritmo poderá apresentar resultados mais precisos na avaliação da defasagem e no cálculo da SNRSEG-NCCF.

5.2.4.2.2 Avaliação da defasagem entre os sinais residuais da fala para o sistema de análise – síntese WI (padrão)

(a) Resultados da aplicação do algoritmo SNRSEG-NCCF para os sinais residuais da fala dos arquivos *casa2*_res_pd.wav e *casa2*_res_rec_pd.wav

A seguir são apresentados os resultados da aplicação do algoritmo SNRSEG-NCCF para a expressão da fala "casa" do locutor – adulto utilizando os sinais residuais da fala armazenados

nos arquivos *casa2*_res_pd.wav (sinal original - referência) e *casa2*_res_rec_pd.wav (sinal reconstruído - em teste). Os resultados são apresentados através dos gráficos nas Figuras 5.19, 5.20, 5.21, 5.22, 5.23 e 5.24.

As Figuras 5.19 e 5.20 mostram os deslocamentos, d' (em amostras), que foram obtidos para cada segmento do arquivo em teste com a aplicação do algoritmo da SNRSEG-NCCF, para segmentos com M = 128 e M = 99 amostras respectivamente.

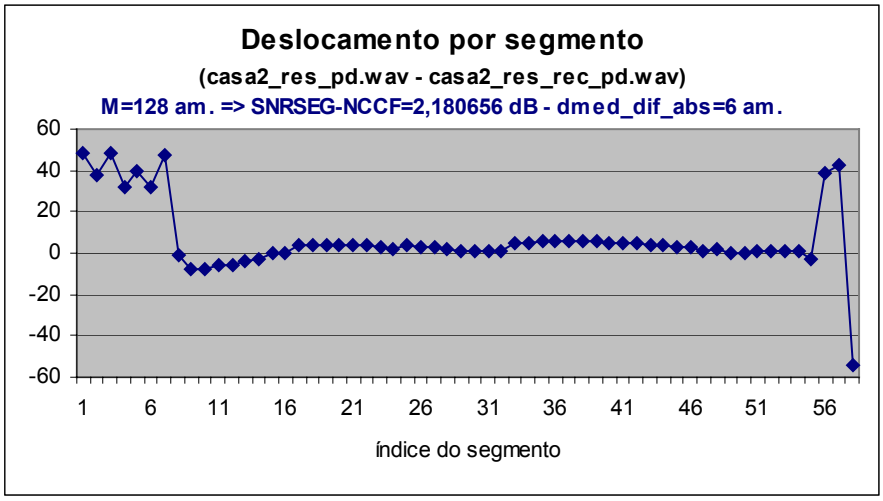

Figura 5.19 – Deslocamentos em amostras, d', para os segmentos do sinal reconstruído, *casa2*_res_rec_pd.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original, *casa2*_res_pd.wav. Resultados obtidos para M =128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

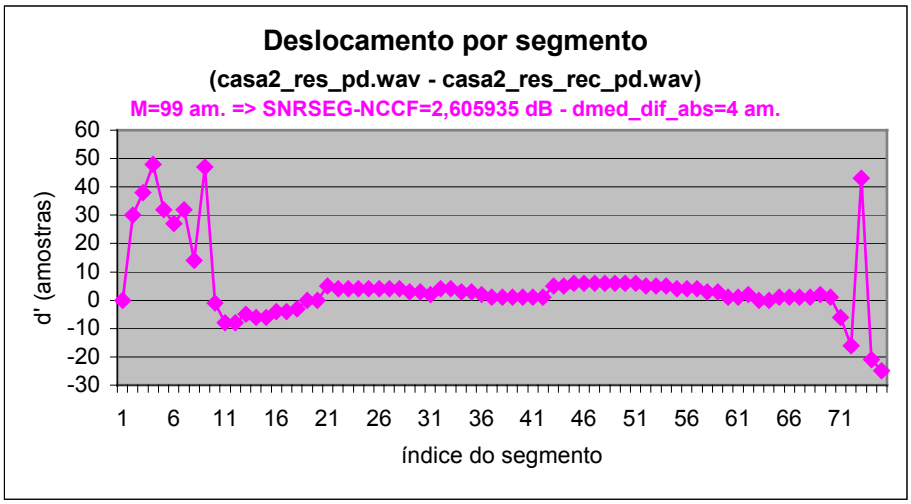

Figura 5.20 – Deslocamentos em amostras, d', para os segmentos do sinal reconstruído, *casa2*_res_rec_pd.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original, *casa2*_res_pd.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

As Figuras 5.19 e 5.20 mostram deslocamentos menores nas regiões sonoras com maior potência. Nas regiões sonoras com potências menores, ou sinais mais ruidosos, os deslocamentos são maiores. Quanto menor for a diferença entre deslocamentos para os segmentos consecutivos menor será a diferença de fase entre os sinais. Observa-se que existem regiões onde os deslocamentos são constantes e diferentes de zero e regiões onde os deslocamentos são constantes e próximo de zero. Pode-se dizer que os sinais estão quase em fase em cada uma dessas regiões. Considerando o sinal completo pode-se dizer que nos trechos inicial e final a fase é variável e que ao longo do trecho central, na maior parte do sinal, a fase é quase constante.

As Figuras 5.21 e 5.22 mostram as correlações cruzadas normalizadas, R (NCCF), entre os segmentos no arquivo *casa2*_res_pd.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_res_rec_pd.wav calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 128 e M = 99 amostras, respectivamente.

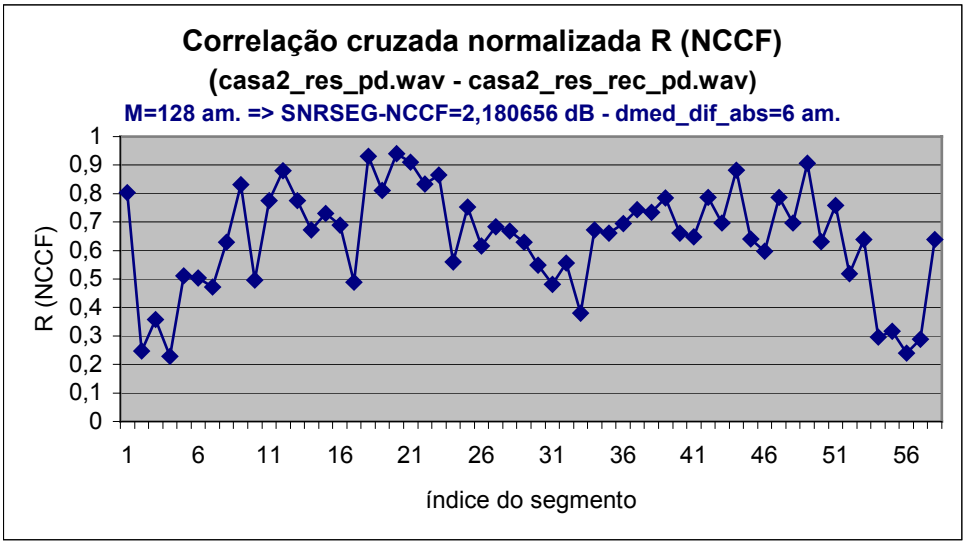

Figura 5.21 – Correlação cruzada normalizada, R (NCCF), entre os segmentos no arquivo *casa2*_res_pd.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_res_rec_pd.wav. Resultados obtidos para M = 128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

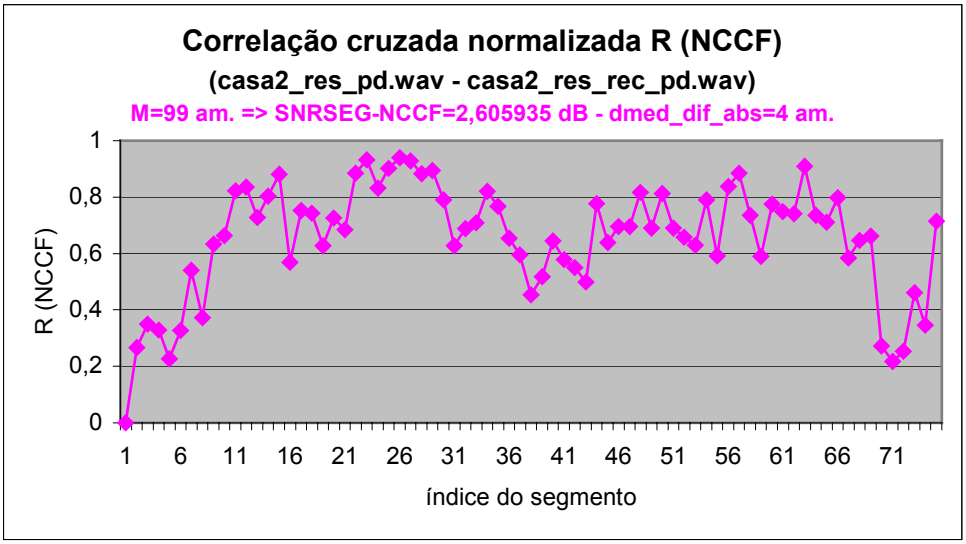

Figura 5.22 – Correlação cruzada normalizada, R (NCCF), entre os segmentos no arquivo *casa2*_res_pd.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_res_rec_pd.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

As Figuras 5.21 e 5.22 mostram que as correlações cruzadas normalizadas são maiores (próximas do valor 1,0) nas regiões onde o sinal é mais potente e o pitch bem definido, ou seja, nas regiões correspondentes onde o sinal da fala é sonoro. Neste caso ao contrário da correlação entre os sinais da fala, Figuras 5.6 e 5.7, existe uma maior variação da correlação R entre os segmentos consecutivos, mostrando uma descontinuidade entre os valores. Os sinais residuais da fala têm faixa dinâmica menor do que a faixa dinâmica dos sinais da fala e são mais ruidosos, o que é indicado pela correlação R menor e mais variável.

Nos extremos onde o sinal é mais ruidoso e com uma potência menor as correlações cruzadas normalizadas são menores.

Pode existir também para o primeiro e último segmento o efeito de borda devido ao algoritmo da SNRSEG-NCCF, quando a janela deslizante enquadra M+M/2 ou M/2+M amostras e as outras M/2 posições da janela são completadas com valores iguais a zero. Assim os valores das correlações cruzadas normalizadas nos extremos dos sinais podem ser desconsiderados.

As Figuras 5.23 e 5.24 mostram as relações sinal ruído, SNR (dB), entre os segmento no arquivo *casa2*_res_pd.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_res_rec_pd.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 128 e M = 99 amostras, respectivamente.

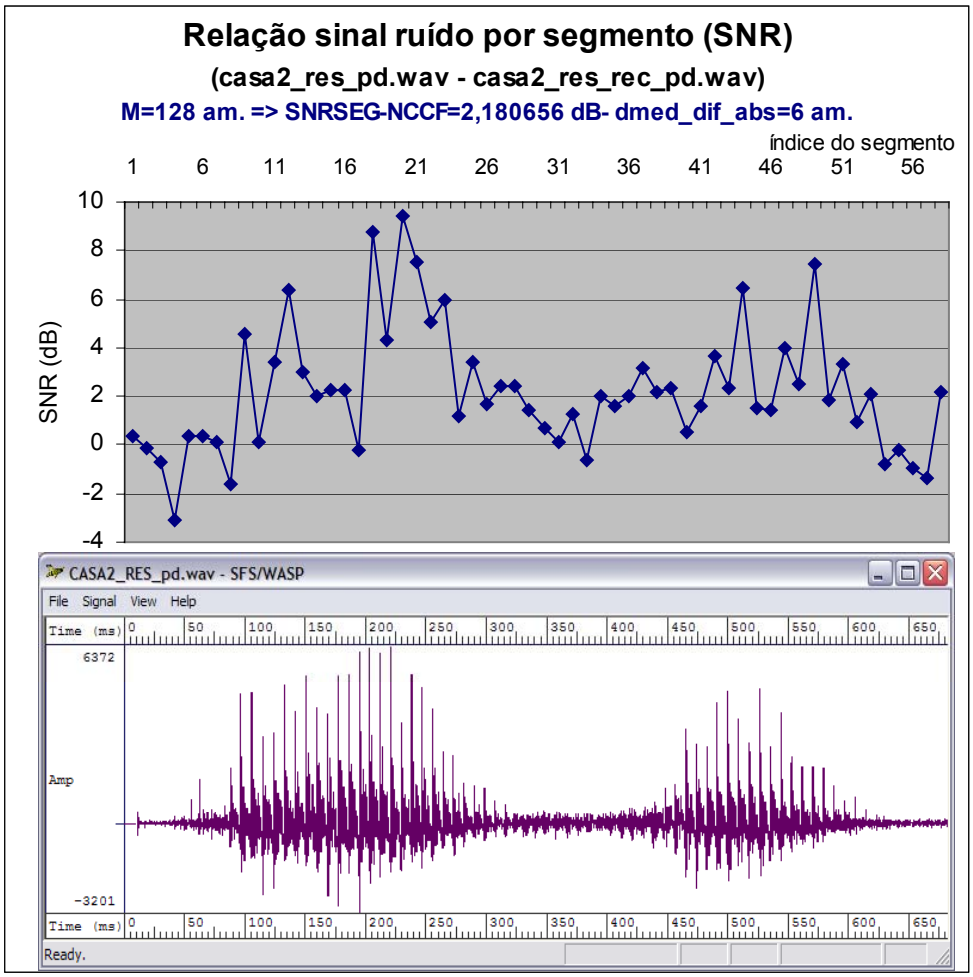

Figura 5.23 – Relação sinal ruído, SNR (dB), entre os segmento no arquivo *casa2*_res_pd.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_res_rec_pd.wav. Para a visualização da posição relativa dos segmentos é mostrado também o gráfico das formas de onda amplitude x tempo para o sinal *casa2*_res_pd.wav. Estes resultados foram obtidos para M = 128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

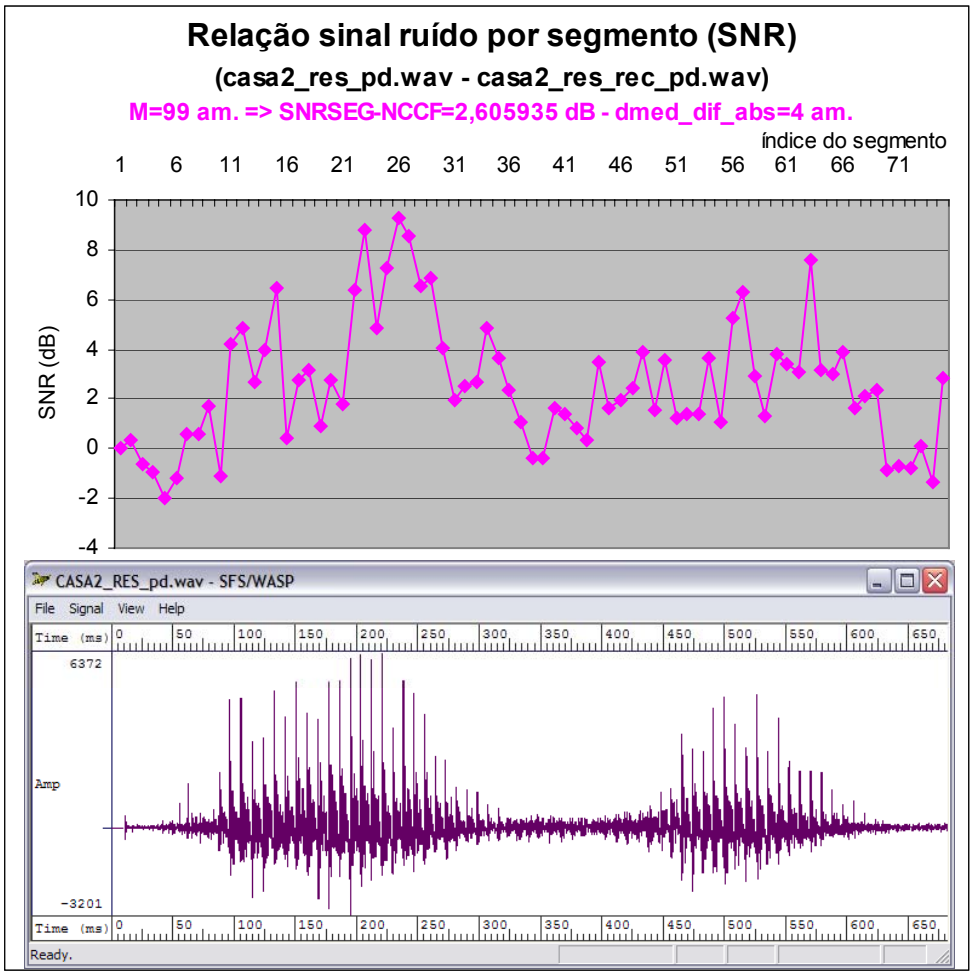

Figura 5.24 – Relação sinal ruído, SNR (dB), entre os segmento no arquivo *casa2*_res_pd.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_res_rec_pd.wav. Para a visualização da posição relativa dos segmentos é mostrado também o gráfico das formas de onda amplitude x tempo para o sinal *casa2*_res_pd.wav. Estes resultados foram obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

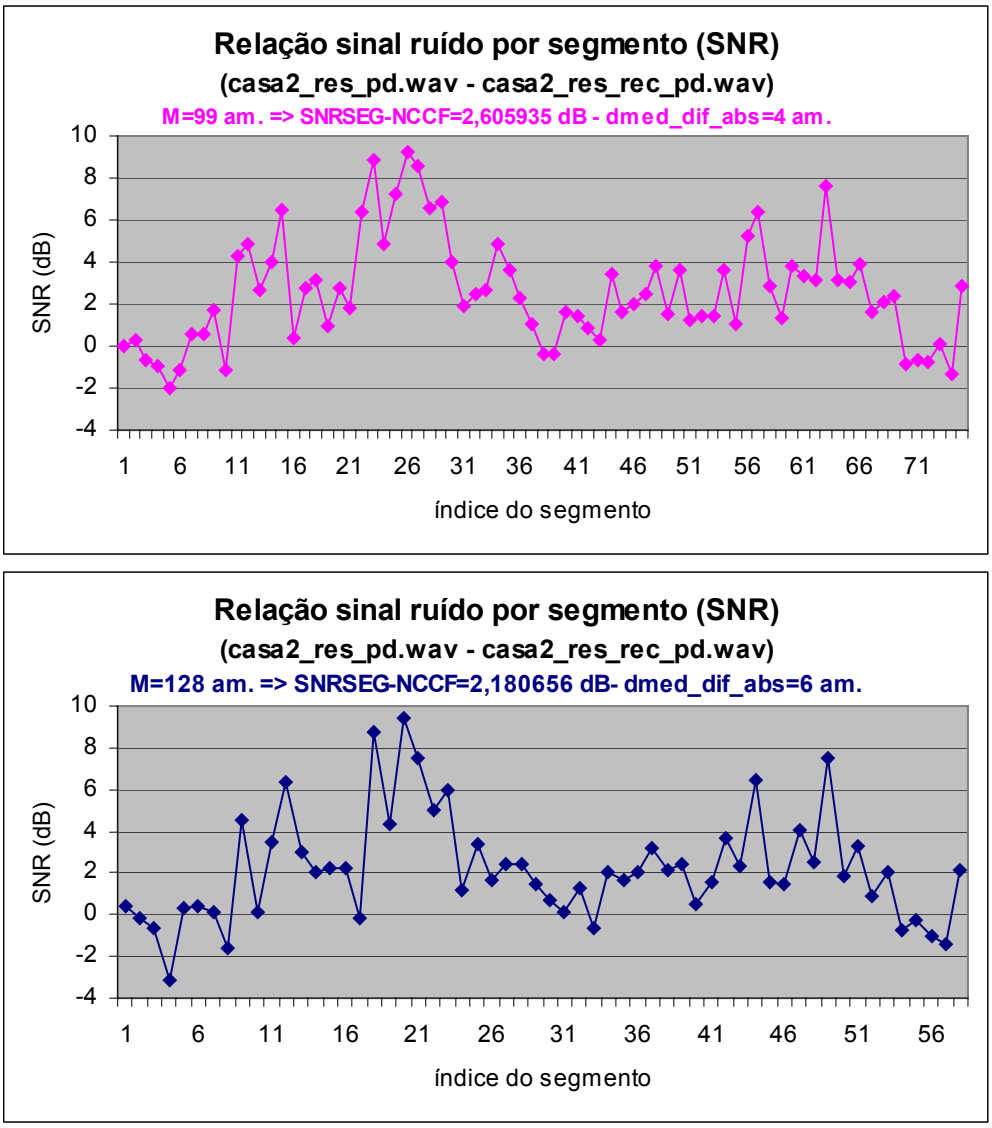

Figura 5.25 - Repetição das Figuras 5.23 e 5.24 para comparação entre o gráfico das SNR's entre os segmentos para M = 128 no gráfico superior e M = 99 amostras no gráfico inferior.

As Figuras 5.23 e 5.24, repetidas na Figura 25, mostram que a SNR por segmento para os sinais residuais também é maior nas regiões correspondentes onde o sinal da fala é sonoro e com maior potência. Mas devido à correlação R ser menor e mais variável que a correlação para os sinais da fala, também de forma geral a SNR por segmento é menor do que no caso dos sinais da fala, como mostrado nas Figuras 5.21 e 5.22. Para os segmentos com M = 99 amostras a medida SNRSEG-NCCF = 2,605935 dB apresentou resultado melhor do que os segmentos com M = 128 amostras onde SNRSEG-NCCF = 2,180656 dB.

(b) Resultados da aplicação do algoritmo SNRSEG-NCCF para os trechos dos sinais residuais sonoros da fala extraídos dos arquivos *casa2*_res_pd.wav e casa2_res_rec_pd.wav

A seguir são apresentados os resultados da aplicação do algoritmo SNRSEG-NCCF para a expressão da fala "casa" locutor – adulto utilizando os trechos com segmentos sonoros dos sinais residuais da fala extraídos dos arquivos *casa2*_res_pd.wav (sinal original - referência) e *casa2*_res_rec_pd.wav (sinal reconstruído - em teste). Os resultados são apresentados em gráficos nas Figuras 5.27, 5.28, 5.29, 5.30 e 5.31 e 5.32.

A Figura 5.26 mostra a localização dos segmentos sonoros dos sinais residuais da fala considerados nos arquivo *casa2*_res_pd.wav e *casa2*_res_rec_pd.wav. Os trechos com sinais sonoros residuais da fala foram extraídos em posições correspondentes nos dois arquivos.

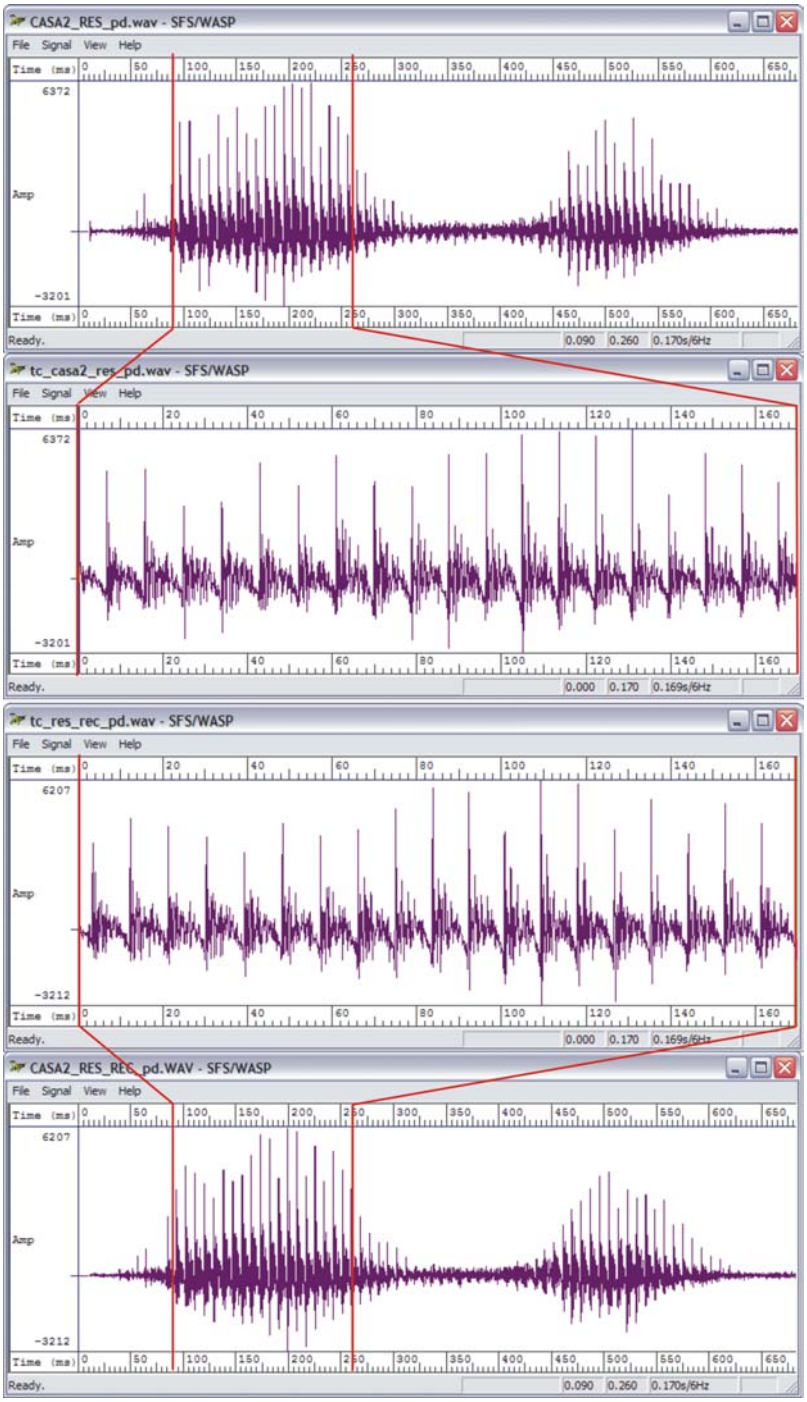

Figura 5.26 - A parte superior da Figura mostra o sinal residual da fala original no arquivo *casa2*_res_pd.wav. Na parte inferior mostra o sinal residual da fala reconstruído no arquivo *casa2*_res_rec_pd.wav. A parte intermediária mostra as ampliações correspondentes aos trechos sonoros extraídos dos sinais residuais da fala, sinal original e sinal reconstruído.

As Figuras 5.27 e 5.28 mostram os deslocamentos, d' (em amostras), que foram obtidos para cada segmento do arquivo em teste (trecho sonoro em *casa2*_res_rec_pd.wav) em relação aos segmentos do arquivo original (trecho sonoro em *casa2*_res_pd.wav) com a

aplicação do algoritmo da SNRSEG-NCCF, para segmentos com M = 128 e M = 99 amostras, respectivamente.

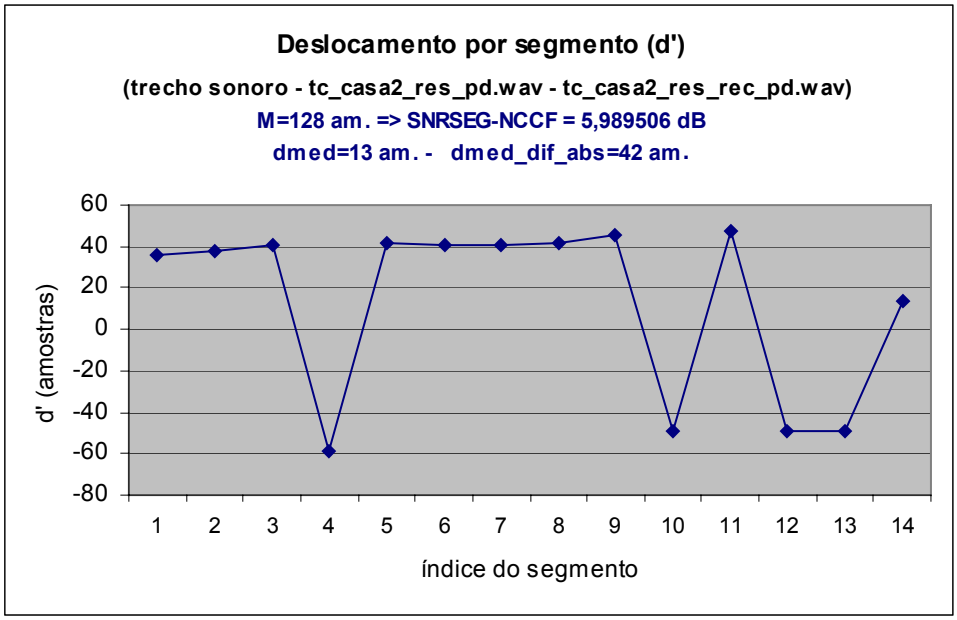

Figura 5.27 – Deslocamentos em amostras, d', para os segmentos do trecho sonoro extraído do sinal residual reconstruído, *casa2*_res_rec_pd.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no trecho sonoro correspondente extraído do sinal residual original, *casa2*_res_pd.wav para M = 128 amostras.

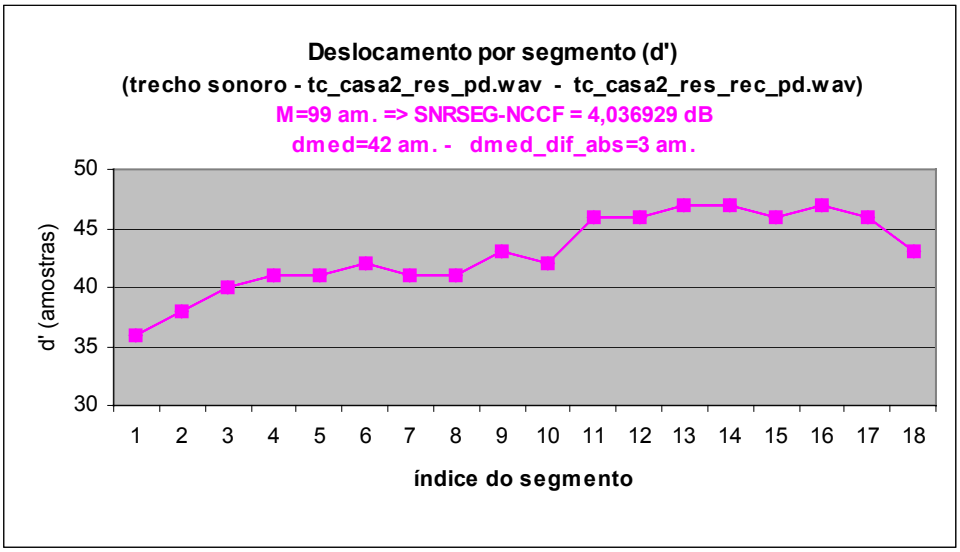

Figura 5.28 – Deslocamentos em amostras, d', para os segmentos do trecho sonoro extraído do sinal residual reconstruído, *casa2*_res_rec_pd.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no trecho sonoro correspondente extraído do sinal residual original, *casa2*_res_pd.wav para M = 99 amostras.

A Figura 5.27 mostra a descontinuidade na seqüência dos segmentos sucessivos. A Figura 5.28 mostra uma seqüência contínua para os segmentos sucessivos. Isto mostra que os segmentos com M = 99 amostras fazem um melhor enquadramento das formas de onda cíclicas dos sinais. Observa-se que a SNRSEG-NCCF para M = 128 amostras foi maior neste caso, trecho no sinal residual correspondente ao trecho sonoro do sinal da fala.

As Figuras 5.29 e 5.30 mostram as correlações cruzadas normalizadas, R (NCCF), entre os segmentos do trecho sonoro extraído do sinal residual *casa2*_res_pd.wav, e os segmentos mais similares próximos às posições correspondentes no trecho sonoro extraído no sinal residual *casa2*_res_rec_pd.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para segmentos com M = 128 e M = 99 amostras, respectivamente.

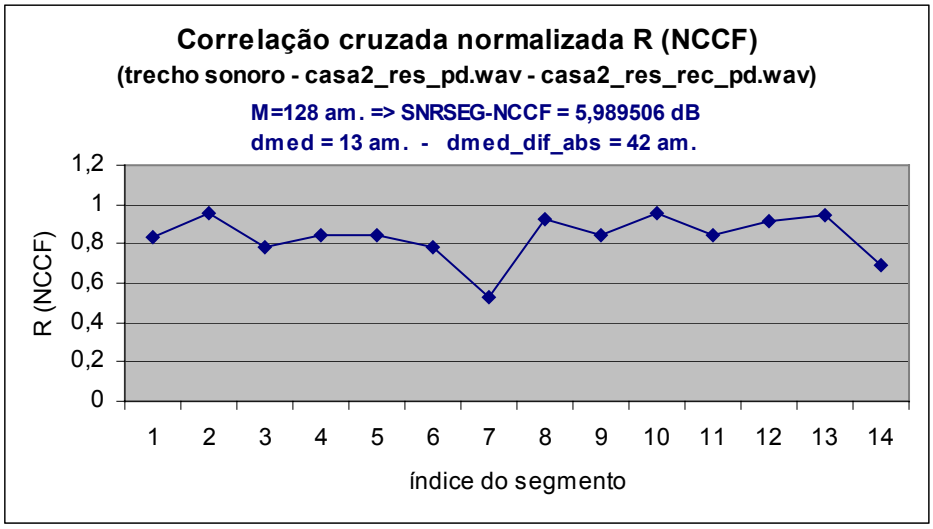

Figura 5.29 – Correlação cruzada, R (NCCF), entre os segmentos do trecho sonoro extraído do sinal residual *casa2*_res_pd.wav, e os segmentos mais similares próximos às posições correspondentes do trecho sonoro extraído do sinal residual *casa2*_res_rec_pd.wav. Resultados obtidos para M = 128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

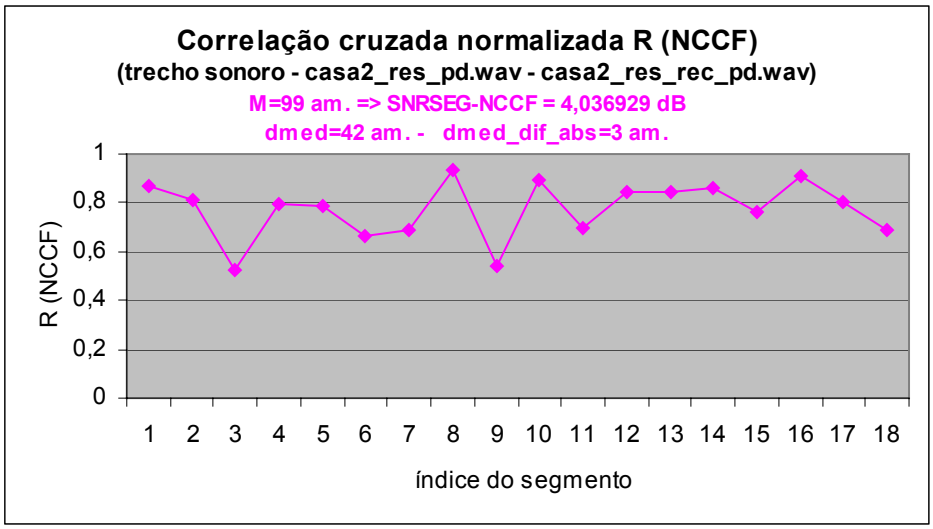

Figura 5.30 – Correlação cruzada, R (NCCF), entre os segmentos do trecho sonoro extraído do sinal residual *casa2*_res_pd.wav, e os segmentos mais similares próximos às posições correspondentes do trecho sonoro extraído do sinal residual *casa2*_res_rec_pd.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

A Figura 5.29 (para M =128 amostras) mostra a correlação R distribuída em uma faixa menor do que na Figura 5.30 (para M = 99 amostras). Comparando as Figuras 5.14 e 5.15 com a correlação no trecho sonoro da fala, observa-se que as correlações R por segmento nos trecho correspondente nos sinais residuais têm uma maior variação entre os segmentos adjacentes e são menores em média.

As Figuras 5.31 e 5.32 mostram as relações sinal ruído, SNR (dB), entre os segmentos do trecho sonoro extraído do sinal residual *casa2*_res_pd.wav e os segmentos correspondentes com a maior similaridade no trecho sonoro extraído do sinal residual *casa2*_res_rec_pd.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 128 e M = 99 amostras, respectivamente.

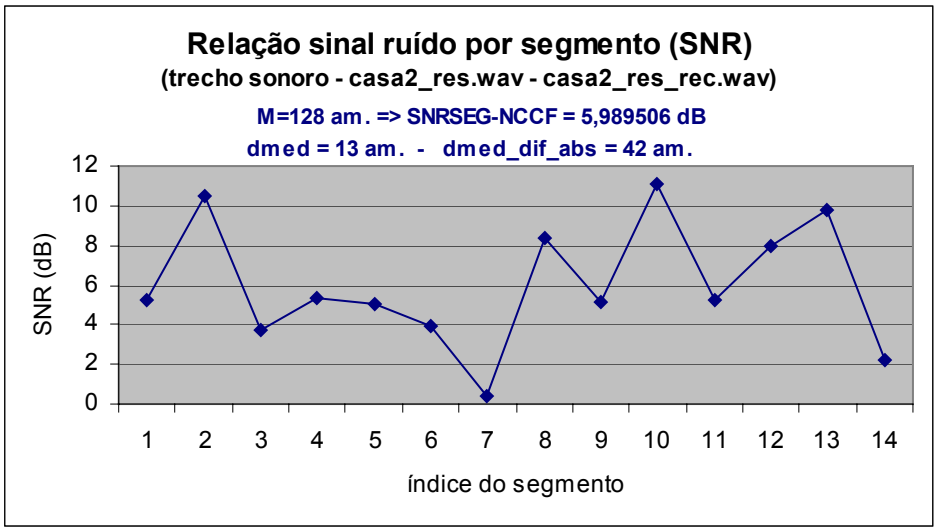

Figura 5.31 – Relação sinal ruído, SNR (dB), entre os segmentos do trecho sonoro extraído do sinal residual *casa2*_res_pd.wav e os segmentos correspondentes com a maior similaridade no trecho sonoro extraído do sinal residual *casa2*_res_rec_pd.wav. Resultados obtidos para M = 128 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

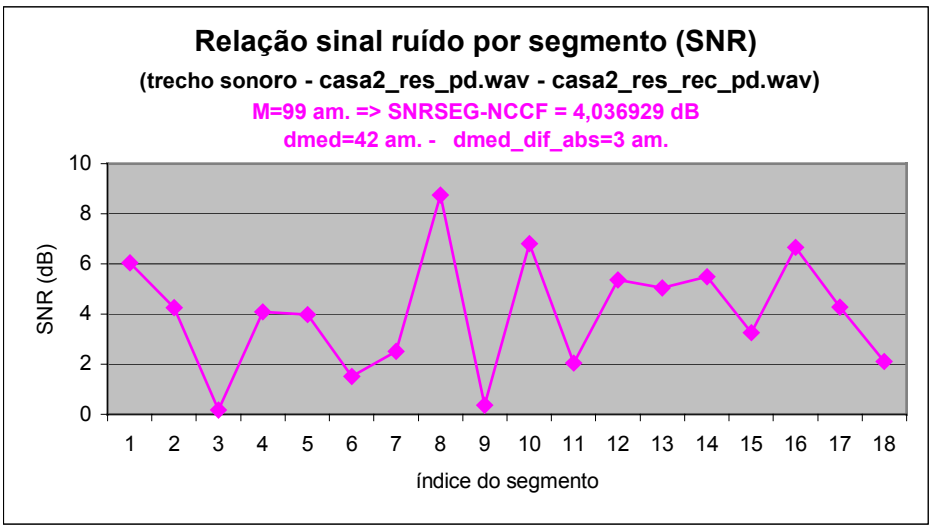

Figura 5.32 – Relação sinal ruído, SNR (dB), entre os segmentos do trecho sonoro extraído do sinal residual *casa2*_res_pd.wav e os segmentos correspondentes com a maior similaridade no trecho sonoro extraído do sinal residual *casa2*_res_rec_pd.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF.

A Figura 5.31 (para M =128 amostras) mostra a uma seqüência de SNR com valores maiores em média que a SNR da Figura 5.32 (para M = 99 amostras) o que é indicado pelas medidas SNRSEG-NCCF = 5,989506 dB (para M =128 amostras) e SNRSEG-NCCF = 4,036929 dB (para M =128 amostras). Ainda que as medidas SNRSEG-NCCF e a correlação R sejam

favoráveis para ao caso de M =128 amostras, os deslocamentos por segmento e as medidas dmed = 42 amostras, dmed_dif_abs = 3 amostras indicam melhores resultados para M = 99 amostras, comparados com dmed = 13 amostras e dmed_dif_abs = 42 amostras dB para M =128 amostras. Isto é devido à medida dmed_dif_abs que indica a diferença média dos deslocamentos entre os segmentos consecutivos. Assim quanto menor dmed_dif_abs menor será a defasagem entre os segmentos e também a dmed = 42 amostras que é um valor mais próximo da defasagem média = 44 amostras indicada no próximo item sobre a inspeção visual.

(c) Inspeção visual para os trechos de sinais sonoros da fala extraídos dos sinais residuais nos arquivos *casa2*_res_pd.wav e casa2_res_rec_pd.wav

A Figura 5.26 mostra os trechos sonoros segmentados dos sinais nos arquivos *casa2*_res_pd.wav e *casa2*_res_rec_pd.wav que foram utilizados para efetuar a análise da defasagem entre os sinais.

Através do software SFS/WASP foram levantados por inspeção direta sobre as formas de ondas dos sinais, o valor do pitch (ciclo a ciclo) e a diferença temporal (a defasagem) entre os picos dos pulsos dos ciclos de pitch, que foram convertidas para número de amostras, considerando a taxa de amostragem de 11025 Hz para os sinais. Os resultados são mostrados na Tabela 5.6.

Tabela 5.6 – Valores obtidos a partir da inspeção visual para os trechos sonoros dos sinais residuais nos arquivos *casa2*_res_pd.wav e *casa2*_res_rec_pd.wav mostrados na Figura 5.26: pitch (Hz) e defasagem (amostras) para freqüência de amostragem de 11025 Hz.

Parâmetros obtidos com a inspeção visual			
Índice dos ciclos do pitch	Trecho sonoro do sinal original *casa2*_res_pd.wav	Trecho sonoro do sinal reconstruído *casa2*_res_rec_pd.wav	Defasagem entre os picos dos pulsos dos ciclos do pitch (amostras)
	Pitch (Hz)	Pitch (Hz)	
1	110	113	44
2	110	113	44
3	113	110	44
4	110	113	44
5	110	110	44
6	113	113	44
7	113	113	44
8	110	113	44
9	116	113	44
10	113	119	33
11	119	119	44
12	113	116	44
13	116	113	44
14	116	116	44
15	116	116	44
16	116	113	55
17	116	116	44
18	113	116	44
19	-	-	44
--	$Pitch_{méd}$ = 113,5 Hz	$Pitch_{méd}$ = 114,17 Hz	$Defasagem_{média}$ = 44 amostras
--	$Pitch_{méd}$= 97,14 amostras	$Pitch_{méd}$ = 96,48 amostras	

A Figura 5.33 mostra o gráfico da defasagem entre os picos dos pulsos de pitch obtidos por inspeção visual de acordo com a Tabela 5.6 para a comparação com o gráfico da Figura 5.28 onde é mostrado o gráfico do deslocamento por segmento para os mesmo trechos do sinal residual,em *casa2*_res_pd.wav e *casa2*_res_rec_pd.wav, que foram obtidos com a aplicação do algoritmo da SNRSEG-NCCF para M = 99 amostras.

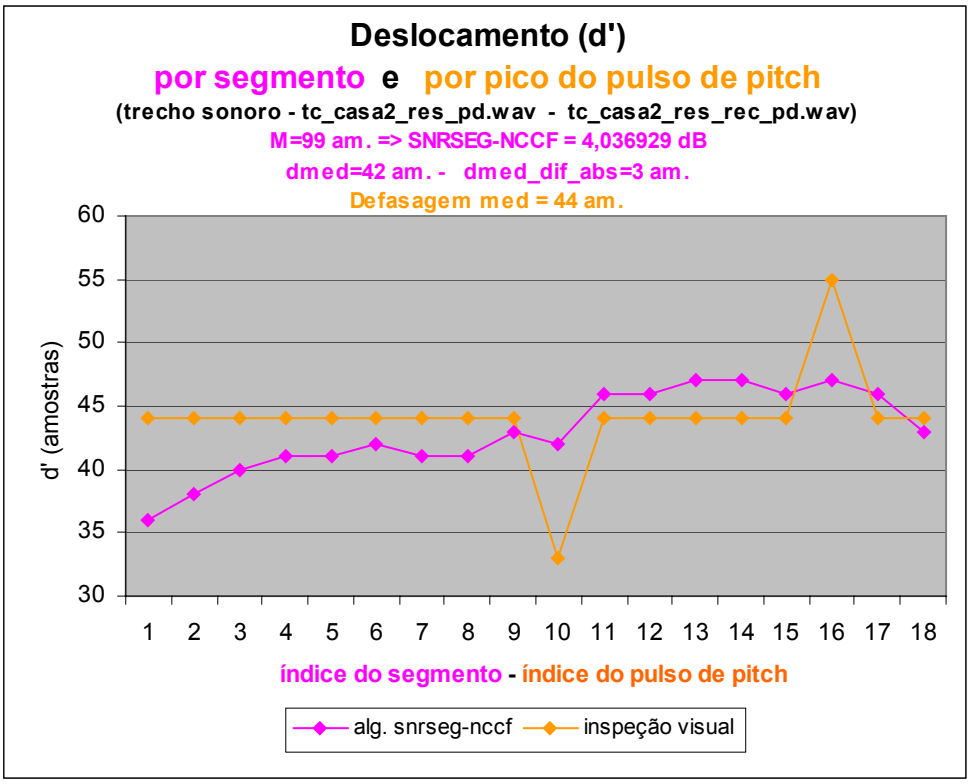

Figura 5.33 - Gráfico para a comparação entre: **(a)** o **deslocamento por segmento** para o trecho do sinal residual com a aplicação do algoritmo da SNRSEG-NCCF (M = 99 amostras) (na cor magenta); e **(b)** a **defasagem entre os picos dos pulsos de pitch** obtidos por inspeção visual de acordo com a Tabela 5.6 (na cor laranja). Estes resultados foram obtidos a partir do trecho extraído do sinal residual correspondente ao trecho sonoro nos sinais residuais reconstruídos no arquivo *casa2*_res_rec_pd.wav e nos sinais residuais originais no arquivo *casa2*_res_pd.wav.

A Figura 5.33 mostra que, para os mesmos trechos *dos sinais residuais* correspondentes aos trechos dos sinais sonoros da fala, os valores para o **deslocamento por segmento** (obtidos *com a aplicação do algoritmo da SNRSEG-NCCF – M = 99 amostras* - na cor magenta) e os valores para a ***defasagem entre os picos dos pulsos de pitch*** *(obtidos por inspeção visual* - na cor laranja) estão aproximadamente dentro de uma faixa semelhante, onde o deslocamento médio foi de 42 amostras e a defasagem media foi de 44 amostras.

(d) Avaliação da defasagem entre os sinais residuais da fala (*casa2*_res_pd.wav *e casa2*_res_rec_pd.wav)

Para avaliação da defasagem entre os sinais residuais da fala, nos arquivos casa2_res_pd.wav e casa2_res_rec_pd.wav, são considerados os resultados obtidos nos itens (a), (b) e (c) que abordam a aplicação do algoritmo SNRSEG-NCCF e a inspeção visual.

Os dados da inspeção visual nos trechos dos sinais residuais correspondentes aos trechos sonoros da fala, mostrados na Tabela 5.6 e na Figura 5.33 (no item (c)), indicam que:

- O pitch manteve-se variando na faixa de 110 a 119 Hz para os sinais de referência e de teste com o $Pitch_{méd}$ = 113,5 Hz (ou 97,14 amostras) para o sinal de referência e o $Pitch_{méd}$ = 114,17 Hz (ou 94,64 amostras) para o sinal em teste;
- Comparando o valor do pitch para as formas de ondas dos ciclos de pitch correspondentes nos dois sinais observa-se coincidência em quase 40% e o restante diferem entre 3 e 6 Hz, o que corresponde a uma variação aproximada em torno do $Pitch_{méd}$ ($Pitch_{méd}$ = 97 amostras) de até 6 amostras aproximadamente para a freqüência de amostragem dos sinais, f_s = 11025 Hz;
- A diferença temporal (ou defasagem em amostras) entre os picos dos pulsos do pitch ficou em média com 44 amostras, variando entre 33 e 55 amostras, sendo que grande parte (89,47 %) permaneceu em torno de 44 amostras, mostrando uma defasagem constante, e a outra parte em torno de 33 amostras (5,26 %) e 55 amostras (5,29 %). Observa-se uma seqüência de segmentos com defasagem constante com 44 amostras e outra seqüência de segmentos com defasagem variável entre 44, 33 e 55 amostras.

Os dados obtidos com a aplicação do algoritmo SNRSEG-NCCF, nos trechos dos sinais residuais correspondentes aos trechos sonoros dos sinais da fala, são mostrados nas Figuras 5.19, 5.20, 5.21, 5.22, 5.23, 5.24, 5.27, 5.28, 5.29, 5.30, 5.31 e 5.32 . Estes dados indicam que:

Para o sinal residual da fala (arquivo completo) - (no item (a)):

- O deslocamento $d_{méd\_dif\_abs}$ foi de 4 amostras para M = 99 amostras e de 6 amostras para M = 128 amostras, indicando que em média os segmentos foram deslocados consecutivamente de 4 amostras e 6 amostras para o cálculo da SNRSEG-NCCF com M = 99 amostras e M = 128 amostras respectivamente;
- A medida SNRSEG-NCCF entre os sinais residuais da fala (referência e reconstruído) foi de 2,605935 dB (para M = 99 amostras) e de 2,180656 dB (para M = 128 amostras);
- A maioria dos resultados, conforme as Tabelas 5.1, 5.2 e 5.3 na Seção 5.2.4.1, também mostra maiores valores para a medida SNRSEG-NCCF e menores valores para $d_{méd\_dif\_abs}$ para M = 99 amostras.

Assim p*ara o sinal residual da fala (arquivo completo)* pode-se dizer que:

- Os sinais apresentaram regiões onde os deslocamentos para os segmentos são variáveis, quase constantes e constantes: *deslocamentos variáveis* – trechos sem sincronismo, correspondentes as posições para os sinais sonoros com baixa potência; *deslocamentos quase constantes* - pouca variação entre os deslocamentos sucessivos, indicando sinais com defasagem quase constante; e *deslocamentos constantes*: pequenos trechos com sinais com

defasagem constante, em posições correspondentes para sinais sonoros com maiores potências..

- De forma geral os sinais residuais também não estão em fase, ou seja, têm fase variável. Para segmentos com M = 99 amostras o algoritmo apresentou os melhores resultados em relação a defasagem. Para as medidas SNRSEG-NCCF o desempenho também foi superior para M = 99 amostras.

Para o trecho residual do sinal da fala correspondente ao trecho sonoro do sinal da fala (no item (b)):

- O deslocamento $d_{méd\_dif\_abs}$ foi de 3 amostras para M = 99 amostras e de 42 amostras para M = 128 amostras, indicando que em média os segmentos foram deslocados consecutivamente de 3 amostras e 42 amostras para o cálculo da SNRSEG-NCCF com M = 99 amostras e M = 128 amostras, respectivamente.
- O deslocamento $d_{méd}$ foi de 42 amostras para M = 99 amostras e de 13 amostras para M = 128 amostras, indicando que em média os segmentos foram deslocados de 42 amostras e 13 amostras para o cálculo da SNRSEG-NCCF com M = 99 amostras e M = 128 amostras, respectivamente.
- A medida SNRSEG-NCCF entre os trechos dos sinais residuais dos sinais da fala (referência e reconstruído) foi de 4,036929 dB (para M = 99 amostras) e de 5,989506 dB (para M = 128 amostras).

Assim para o trecho dos sinais residuais correspondentes ao trecho sonoro do sinal da fala pode-se dizer que:

- O trecho dos sinais residuais correspondentes ao trecho sonoro do sinal da fala confirma os melhores resultados para os segmentos com M = 99 amostras. O algoritmo apresentou os melhores resultados em relação à defasagem $d_{méd\_dif\_abs}$ = 3 amostras comparada com $d_{méd\_dif\_abs}$ = 42 amostras para os segmentos com M = 128 amostras. E também em relação ao deslocamento médio $d_{méd}$ que foi de $d_{méd}$ = 42 amostras para M = 99 amostras, comparado com $d_{méd}$ = 13 amostras para M = 128 amostras. Neste caso $d_{méd}$ = 42 amostras para M = 99 amostras está mais próximo da defasagem média = 44 amostras obtida por inspeção visual. Para as medidas SNRSEG-NCCF o desempenho foi inferior com 4,036929 dB (para M = 99 amostras) e 5,989506 dB (para M = 128 amostras).

Considerações finais (seção 5.2.4.2.2) - Com os dados apresentados para os sinais residuais da fala, nos arquivos casa2_res_pd.wav e casa2_res_rec_pd.wav pode-se dizer que:

- De forma geral os sinais residuais não estão em sincronismo, ou em fase;
- Nos trechos dos sinais residuais correspondentes aos trechos sonoros dos sinais da fala com maiores potências o algoritmo mostra que os sinais apresentam regiões com defasagem quase constante e regiões com defasagem constante entre os segmentos;

- A inspeção visual mostrou que a defasagem entre o trecho do sinal residual original da fala e o trecho do sinal residual reconstruído manteve-se em torno do valor médio de 44 amostras, na faixa de 33 a 55 amostras, mostrando trechos constantes em 44 amostras. De forma global existe uma defasagem constante de 44 amostras em média para o trecho do sinal residual correspondente ao trecho sonoro da fala;
- Nos trechos do sinal residual correspondente ao trecho sonoro da fala com baixas potências, ou sinais mais ruidosos, o algoritmo mostra a falta de sincronismo (defasagem variável);
- O algoritmo apresentou melhores resultados para o trecho do sinal residual correspondente ao trecho sonoro da fala com M = 99 amostras próximo ao valor do pitch médio de 97 amostras. Isto é um indicativo, a ser investigado nas pesquisas futuras, de que se houver sincronismo entre o tamanho M dos segmentos com o valor do pitch médio, o algoritmo poderá apresentar resultados mais precisos na avaliação da defasagem e no cálculo da SNRSEG-NCCF.

5.2.4.2.3 Avaliação geral da defasagem no sistema de análise – síntese WI (padrão): sinais da fala e sinais residuais (*casa2*.wav *e casa2*_sin_rec_pd.wav; *casa2_res*_pd.wav *e casa2*_res_rec_pd.wav)

Na Seção 5.2.4.2 foi proposto avaliar a defasagem (ou diferença de fase) entre os sinais da fala (*originais* – ou sinal de referência e *reconstruídos* - ou sinal em teste) processados no sistema de análise – síntese WI (padrão). Para fazer esta avaliação foram utilizados o algoritmo da SNRSEG-NCCF (nos sinais da fala e nos sinais residuais da fala) e uma inspeção visual das formas de onda (em trecho sonoro dos sinais da fala e dos sinais residuais).

A partir do método da SNRSEG-NCCF foram calculados os parâmetros na procura dos segmentos com maior similaridade: os deslocamentos dos segmentos, **d'**; as correlações cruzadas normalizadas, **Rmáx**; o deslocamento médio **d'méd**; a medida da relação sinal ruído segmental com auxílio da correlação cruzada normalizada, **SNRSEG-NCCF**; e o valor médio das diferenças absolutas para os deslocamentos dos segmentos sucessivos, $\mathbf{d_{méd\_dif\_abs}}$. As correlações **Rmáx** são utilizadas no cálculo do deslocamento **d'**. Os deslocamentos **d'** indicam a desfasagem por segmento, ou seja, fornece uma noção do defasamento do sinal enquadrado em cada segmento. O deslocamento médio **d'méd,** ou defasagem média, foi útil durante o controle do defasamento geral entre os sinais (referência e teste) para a leitura nos dois arquivos, ou seja, no controle do início da leitura dos dois arquivos. Para os arquivos completos, dos sinais da fala ou dos sinais residuais da fala, procurou-se durante as execuções do algoritmo controlar o início da leitura dos arquivos para manter d'méd próximo de zero. Para os trechos dos sinais sonoros ou trechos dos sinais residuais correspondentes, o

deslocamento médio d'méd também foi útil para a comparação com a ***defasagem média*** obtida por inspeção visual na verificação da defasagem entre os sinais. O valor médio das diferenças absolutas para os deslocamentos entre os segmentos sucessivos, $\mathbf{d_{méd\_dif\_abs}}$ indica a defasagem média absoluta entre os segmentos sucessivos. Este parâmetro e a medida SNRSEG-NCCF foram considerados como os principais valores na avaliação da defasagem entre os sinais. Os resultados dos experimentos relatados nos itens anteriores são listados na Tabela 5.7 para uma comparação dos parâmetros obtidos entre os sinais da fala originais e reconstruídos, e sinais residuais da fala originais e reconstruídos.

De acordo com os resultados e as avaliações feitas nos itens anteriores para sinais da fala e sinais residuais da fala, existe diferença de fase nos sinais da fala e nos sinais residuais da fala após o processamento no sistema de análise – síntese WI (padrão). De forma geral, os resultados mostram que a defasagem é variável. Nos trechos sonoros da fala e nos sinais residuais com baixas potências, ou sinais mais ruidosos, o algoritmo mostra que falta sincronismo (a defasagem é variável). Em trechos com sinais sonoros com maior energia os sinais apresentam regiões com defasagem quase constante e regiões com defasagem constante entre os segmentos. Nestes casos o defasamento ficou praticamente constante, ou seja, existe uma diferença fixa ou quase fixa da fase entre os ciclos do pitch correspondentes e consecutivos nos sinais original e reconstruído. Na Tabela 5.7 observa-se que o trecho dos sinais residuais tem $d_{méd\_dif\_abs}$ de 3 amostras para M = 99 amostras e de 42 amostras para M = 128 amostras enquanto que para o trecho dos sinais da fala $d_{méd\_dif\_abs}$ foi de 2 amostras para M = 99 amostras e de 4 amostras para M = 128 amostras. O processamento do algoritmo SNRSEG-NCCF com o tamanho dos segmentos M = 99 amostras, aponta para os menores valores de $d_{méd\_dif\_abs}$ e também, para o caso dos trechos sonoros, para **d'méd** próximo da ***defasagem média*** (obtida por inspeção visual) e portanto para resultados mais confiáveis para a defasagem entre os sinais. No caso analisado observa-se que M é um valor próximo do valor do pitch médio para os trechos dos sinais considerados. Desta forma, estes resultados mostram que para os segmentos com tamanho M próximo ao valor do pitch médio para o trecho sonoro da fala e para o trecho do sinal residual levam a uma avaliação com maior precisão para a defasagem entre os sinais. Isto significa que o segmento faz um melhor enquadramento do ciclo do pitch (ou trecho do ciclo). Assim existem indícios de que se o tamanho do segmento estiver sincronizado com o valor do pitch, o algoritmo SNRSEG-NCCF poderá efetuar uma melhor avaliação da defasagem e do cálculo da SNRSEG-NCCF.

Tabela 5.7 - Resumo dos resultados**:** Avaliação geral da defasagem para sinais da fala e sinais residuais da fala no sistema de análise – síntese WI (padrão).

Parâmetros para a avaliação geral da defasagem na expressão da fala "casa" - locutor adulto : **sinais da fala:** arquivo completo: ***casa2*.wav e *casa2*_sin_rec_pd.wav –** trecho sonoro da fala***: tc_casa2*_**pd.wav e *tc_casa2*_sin_rec_pd.wav** **sinais residuais:** arquivo completo: ***casa2*_res_pd.wav e *casa2*_res_rec_pd.wav** trecho sonoro do sinal residual da fala: ***tc_casa2*_res_pd.wav e *tc_casa2*_res_rec_pd.wav**									
		Sinais da fala				**Sinais residuais da fala**			
		Sinal original	**Sinal Recons-truído**	**Sinal original**	**Sinal Recons-truído**	**Sinal original**	**Sinal Recons-truído**	**Sinal original**	**Sinal Recons-truído**
		arquivo completo		*Trecho sonoro*		*arquivo completo*		*Trecho sonoro*	
Pitch médio (Hz) (Ins.Vis.)		---	---	114,27	114,27	---	---	113,5	114.17
(Inspeção visual) **Faixa de variação do pitch (Hz)**		---	---	108 a 120	111 a 117	---	---	110 a 119	110 a 119
Deslocamento Médio - d'$_{méd}$ (amostras)	M = 99 amostras	__		26		__		42	
	M = 128 amostras	__		25		__		13	
(Inspeção visual) **Defasagem média (amostras)**		__	__	29		__	__	44	
Média das diferenças absolutas dos deslocamentos entre os segmentos $d_{méd\_dif\_abs}$ (amostras)	M = 99 amostras	6		2		4		3	
	M = 128 amostras	9		4		6		42	
Relação sinal ruído segmental com NCCF (SNRSEG-NCCF) (dB)	M = 99 amostras	11,959578		12,403888		2,605935		4,036929	
	M = 128 amostras	11,378476		11,842292		2,180656		5,989506	

5.2.4.3 Os resultados: gráficos das formas de onda para o sistema de análise – síntese WI (padrão)

Nesta seção são mostrados os gráficos das formas de onda (amplitude *versus* tempo) para a comparação visual de alguns sinais da fala originais, sinais da fala reconstruídos, sinais residuais da fala (na análise) e sinais residuais reconstruídos (na síntese).

5.2.4.3.1 Gráficos das formas de onda (locutora - adulta) - sistema de análise – síntese WI (padrão)

I - Expressão da fala (locutora - adulta): "**casa**"

(a) A Figura 5.34 mostra o sinal da fala original *casa1*.wav, o sinal da fala reconstruído ***casa1*_sin_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais original e reconstruído pelo sistema de análise – síntese WI (padrão).

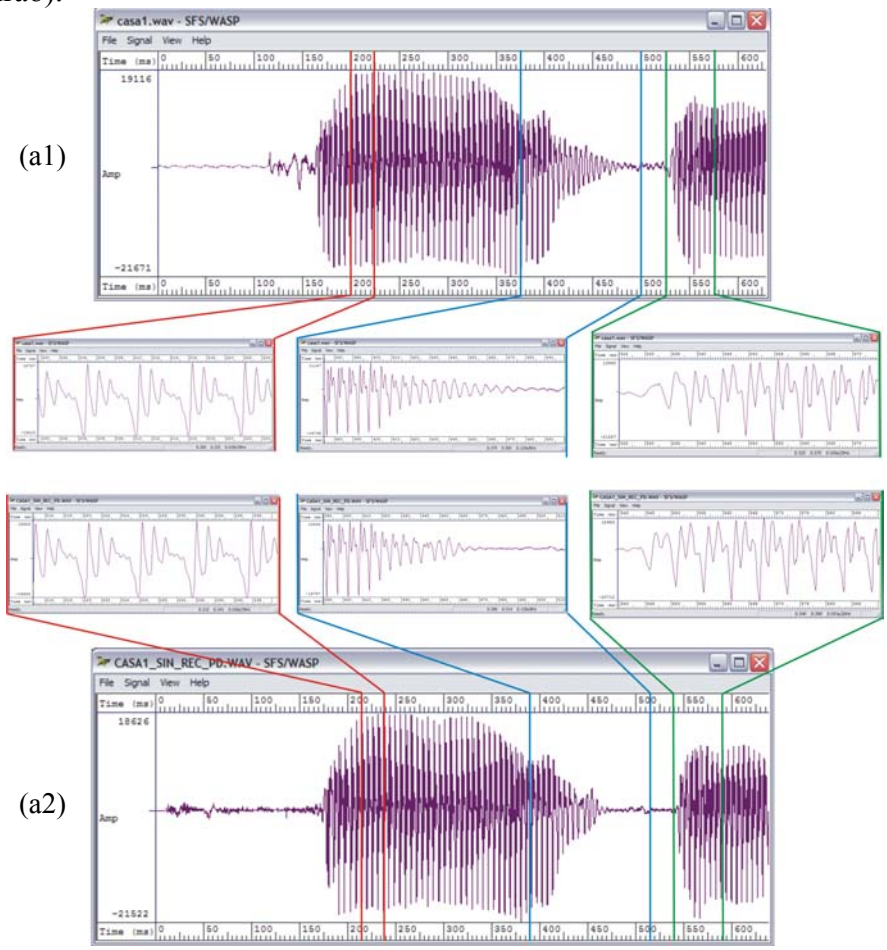

Figura 5.34 – Gráfico das formas de onda (amplitude x tempo) para os sinais (locutora adulta): (a1) Sinal da fala original: **casa1.wav;** (a2) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): **casa1_sin_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (a1) e (a2).

Na Figura 5.34, comparando os gráficos (a1) e (a2), observa-se que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados também mostram grande similaridade através das formas de onda principalmente nas regiões sonoras. Na região inicial correspondente ao fonema /c/ nota-se diferença entre as formas de ondas.

(b) A Figura 5.35 mostra o sinal residual original da fala do arquivo ***casa1*_res_pd.wav**, o sinal residual reconstruído da fala do arquivo ***casa1*_res_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais residuais original e o reconstruído pelo sistema WI (padrão).

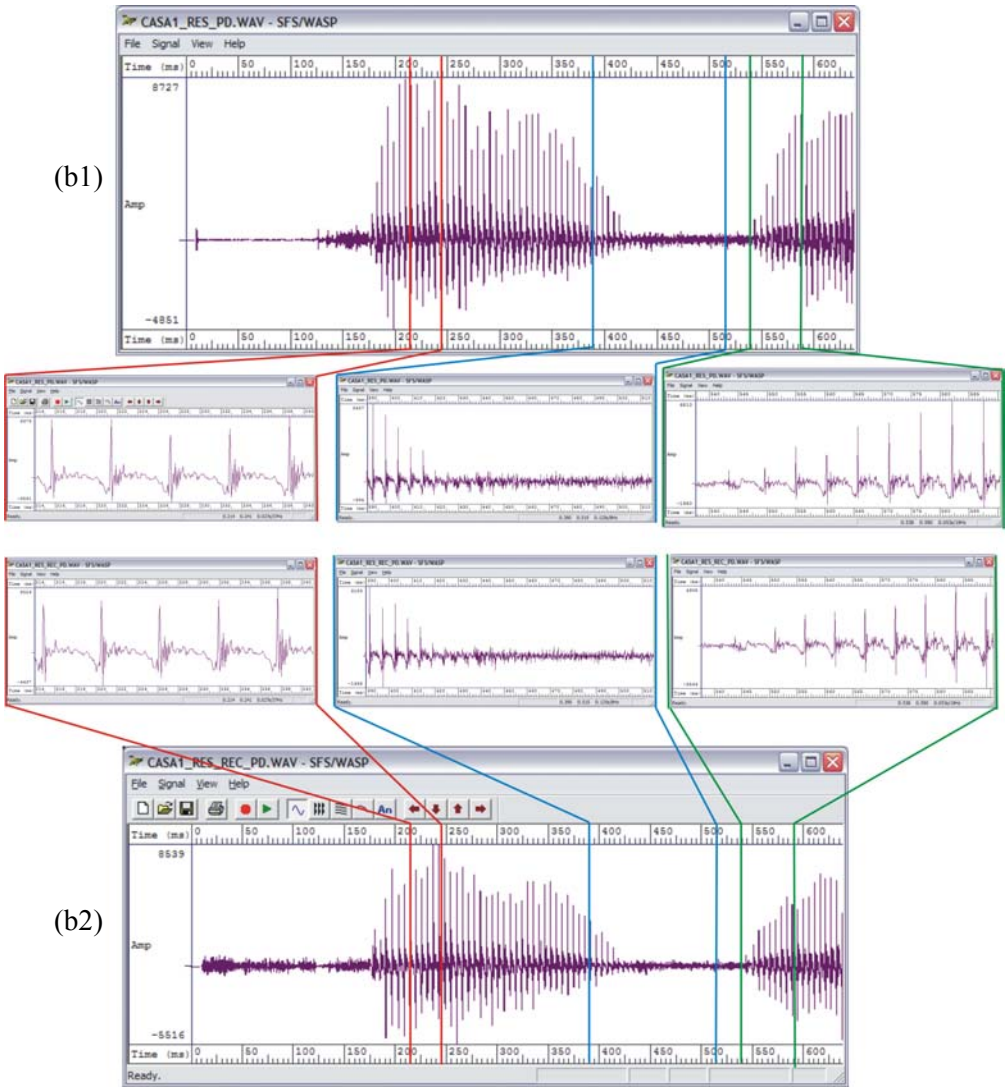

Figura 5.35 – Gráfico das formas de onda (amplitude x tempo) para os sinais residuais (locutora adulta): (b1) Sinal residual original da fala: do arquivo **casa1_res_pd.wav;** (b2) Sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo **casa1_res_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (b1) e (b2).

Na Figura 5.35, comparando os gráficos (b1) e (b2), observa-se que existem semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados, correspondentes aos segmentos sonoros da fala, também mostram grande similaridade através das formas de ondas. Na região inicial mais ruidosa, correspondente ao fonema /c/, nota-se diferença entre as formas de ondas.

II - Expressão de fala (locutora - adulta): "**é bonita**"

(a) A Figura 5.36 mostra o sinal da fala original do arquivo ***ebonita1*.wav**, o sinal da fala reconstruído do arquivo ***ebonita1*_sin_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais original e reconstruído pelo sistema WI (padrão).

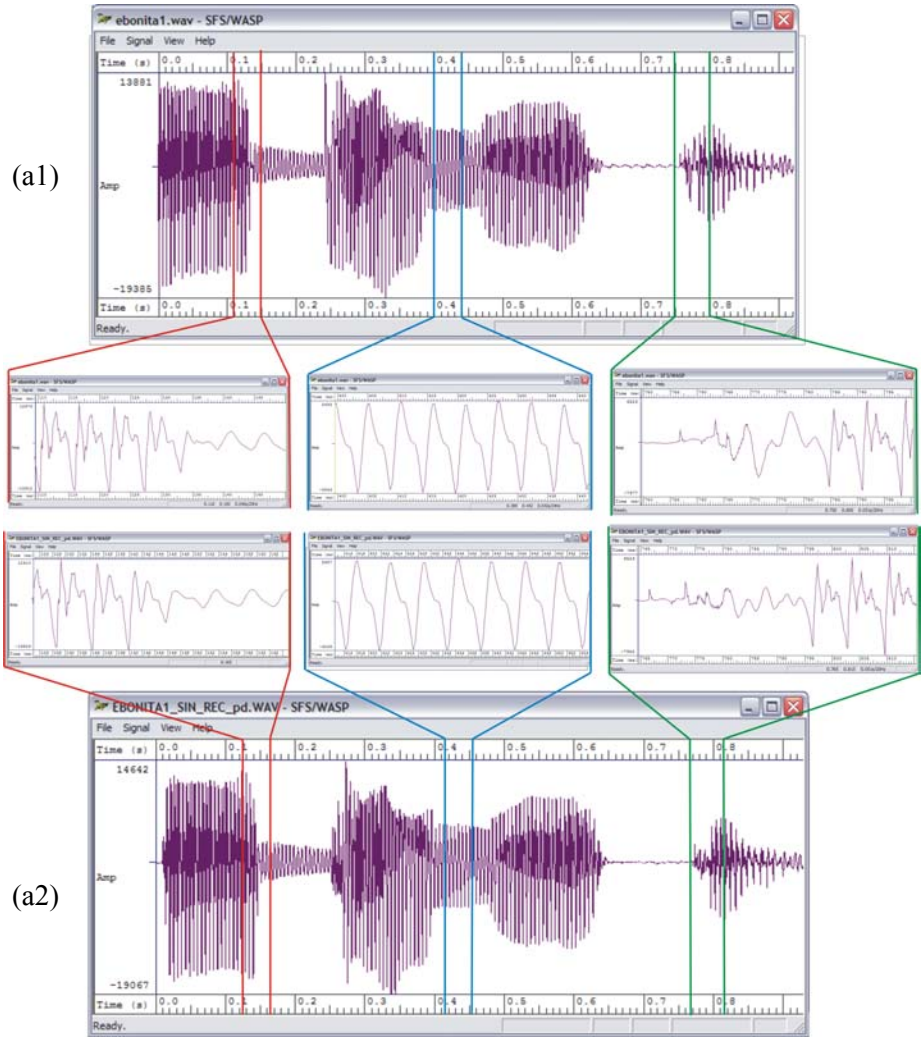

Figura 5.36 – Gráfico das formas de onda (amplitude x tempo) para os sinais (locutora adulta): (a1) Sinal da fala original: do arquivo **ebonita1.wav;** (a2) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo **ebonita1_sin_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (a1) e (a2).

Na Figura 5.36, comparando os gráficos (a1) e (a2), observa-se que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados também mostram grande similaridade através das formas de onda principalmente nas regiões sonoras. No último trecho ampliado, na região final do sinal, correspondente à transição entre os fonemas /t/ e /a/ nota-se a diferença entre as amplitudes das formas de ondas.

(b) A Figura 5.37 mostra o sinal residual original da fala do arquivo ***ebonita1_res_pd*.wav**, o sinal residual reconstruído da fala do arquivo ***ebonita1*_res_rec_pd.wav** para e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais residuais original e o reconstruído pelo sistema WI (padrão).

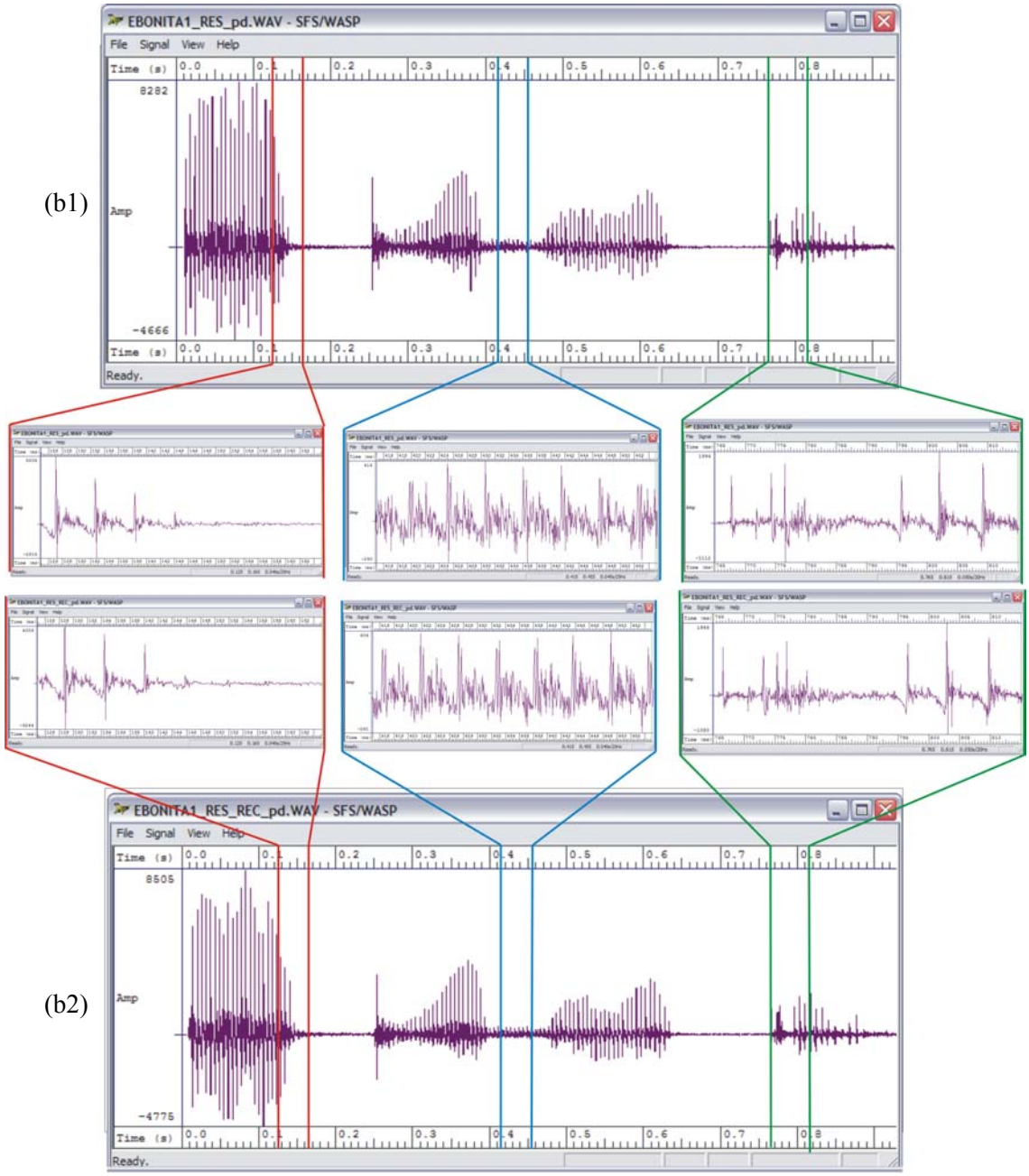

Figura 5.37 – Gráfico das formas de onda (amplitude x tempo) para os sinais residuais (locutora adulta): (b1) Sinal residual original da fala: do arquivo ***ebonita1*_res_pd.wav;** (b2) Sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo ***ebonita1*_res_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (b1) e (b2).

Na Figura 5.37, comparando os gráficos (b1) e (b2), observa-se que existem semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados, correspondentes aos segmentos sonoros da fala, também mostram grande similaridade através das formas de ondas.

5.2.4.3.2 Gráficos das formas de onda (locutor - adulto) - sistema de análise – síntese WI (padrão)

I - Expressão de fala: "**casa**" (locutor - adulto)

(a) A Figura 5.38 mostra o sinal da fala original do arquivo ***casa2*.wav**, o sinal da fala reconstruído do arquivo ***casa2*_sin_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais original e reconstruído pelo sistema WI (padrão).

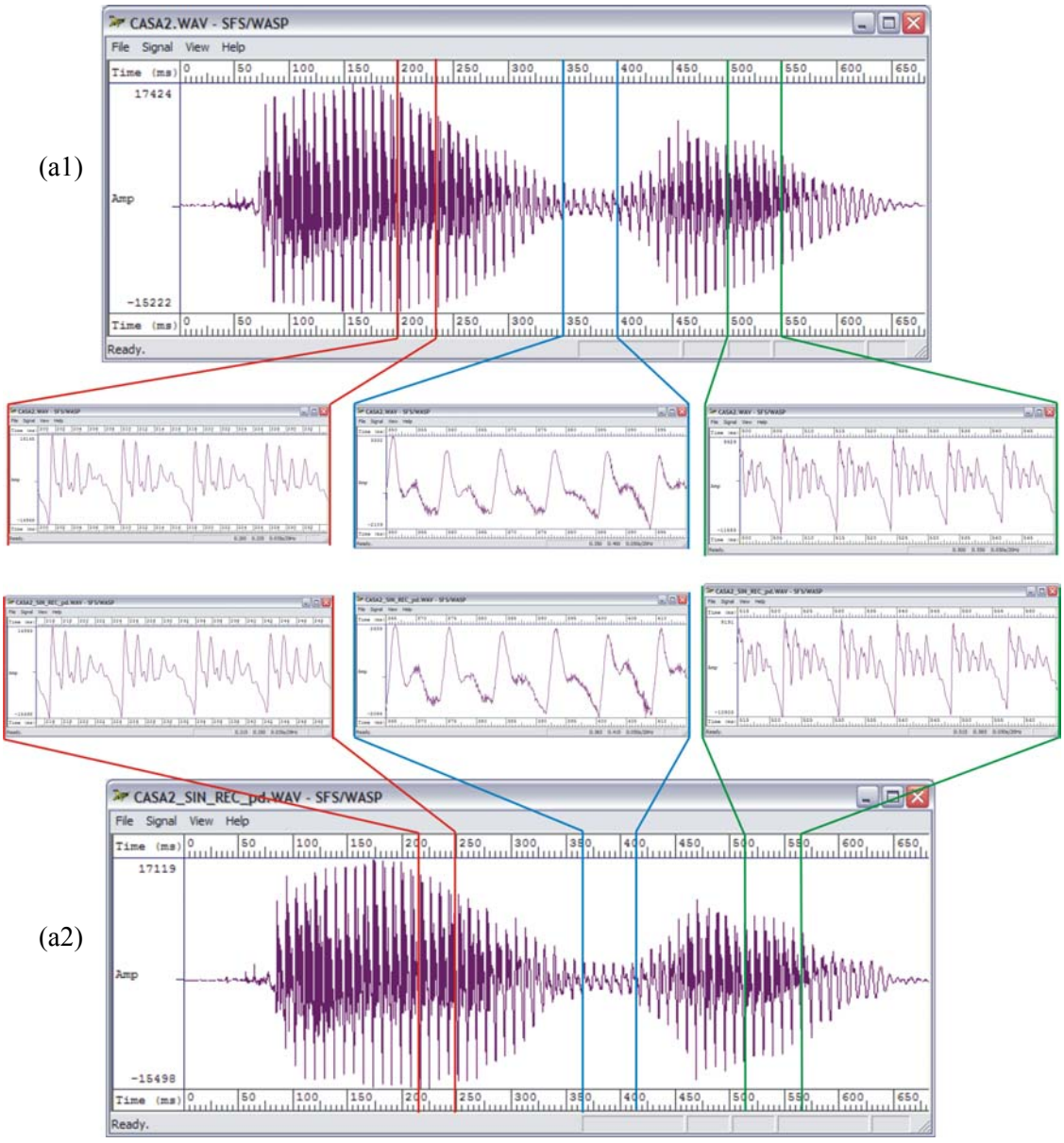

Figura 5.38– Gráfico das formas de onda (amplitude x tempo) para os sinais (locutor adulto): (a1) Sinal da fala original: do arquivo **casa2.wav;** (a2) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo **casa2_sin_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (a1) e (a2).

Na Figura 5.38, comparando os gráficos (a1) e (a2), observa-se que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados também mostram grande similaridade através das formas de onda principalmente nas regiões sonoras.

(b) A Figura 5.39 mostra o sinal residual original da fala do arquivo ***casa2_res_pd*.wav**, o sinal residual reconstruído da fala do arquivo ***casa2*_res_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais residuais original e o reconstruído pelo sistema WI (padrão).

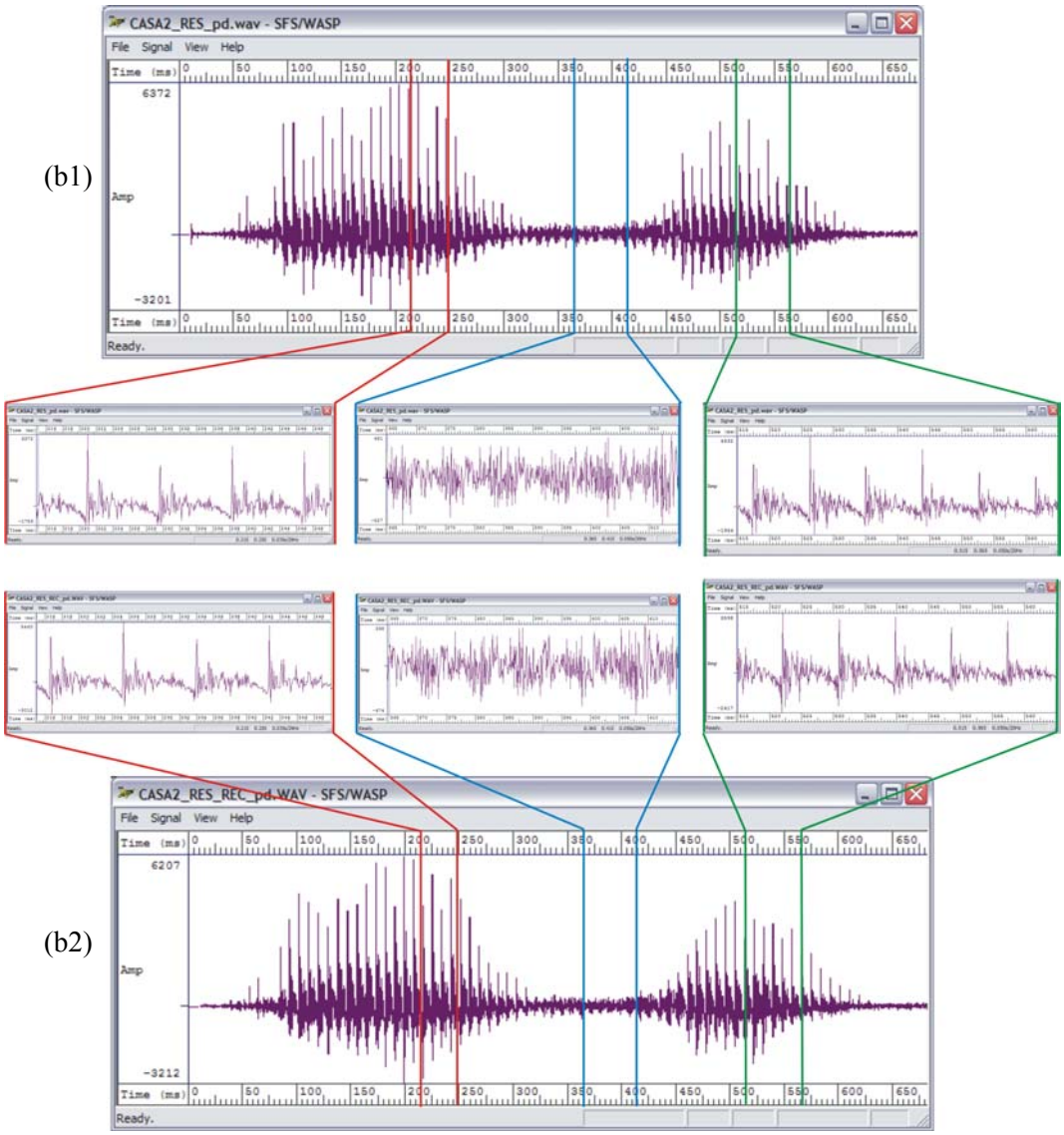

Figura 5.39 – Gráfico das formas de onda (amplitude x tempo) para os sinais residuais (locutor adulto): (b1) Sinal residual original da fala: do arquivo **casa2_res_pd.wav;** (b2) Sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo **casa2_res_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (b1) e (b2).

Na Figura 5.39, comparando os gráficos (b1) e (b2), observa-se que existem semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados, correspondentes aos segmentos sonoros da fala, também mostram grande similaridade nas formas de ondas.

II - Expressão de fala: "**é bonita**" (locutor adulto)

(a) A Figura 5.40 mostra o sinal da fala original do arquivo ***ebonita2*.wav**, o sinal da fala reconstruído do arquivo ***ebonita2*_sin_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais original e reconstruído pelo sistema WI (padrão).

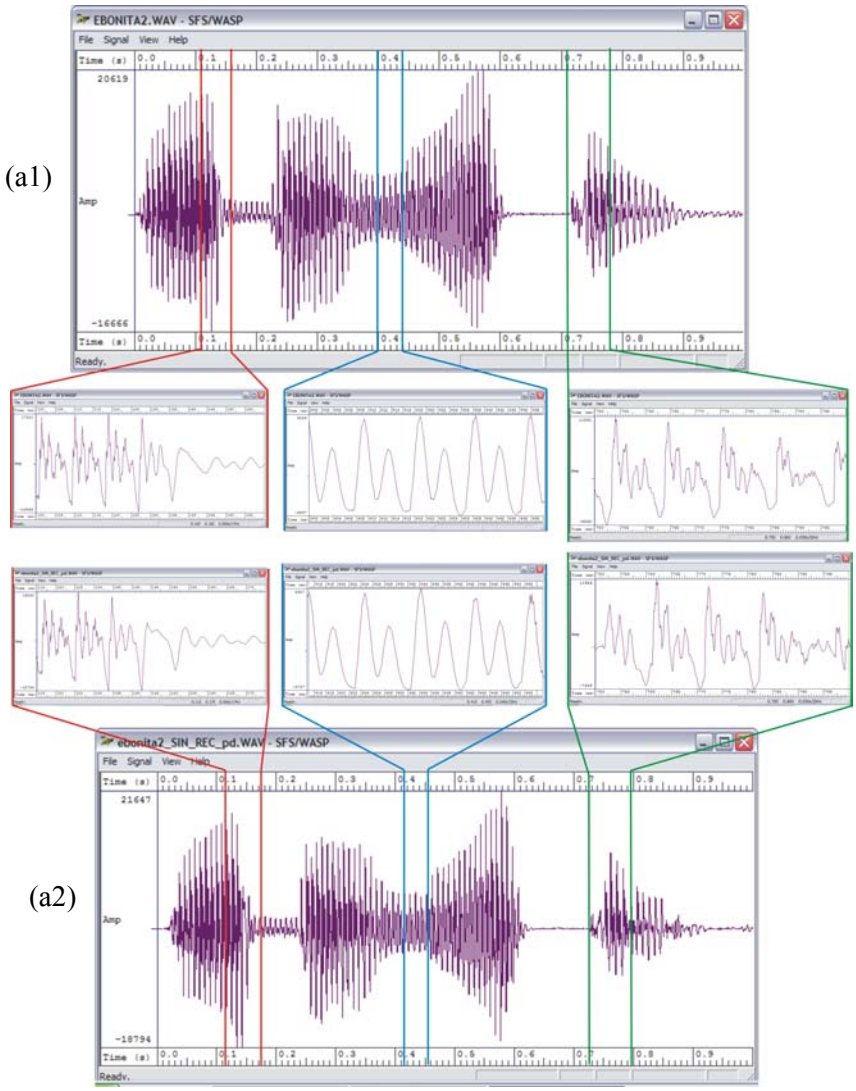

Figura 5.40– Gráfico das formas de onda (amplitude x tempo) para os sinais (locutor adulto): (a1) Sinal da fala original: do arquivo **ebonita2.wav;** (a2) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo **ebonita2_sin_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (a1) e (a2).

Na Figura 5.40, comparando os gráficos (a1) e (a2), observa-se que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados também mostram grande similaridade nas formas de ondas principalmente nas regiões sonoras. No último trecho ampliado e fora do trecho, na região final do sinal, correspondente ao fonema /a/ nota-se diferença entre as amplitudes das formas de ondas.

(b) A Figura 5.41 mostra o sinal residual original da fala do arquivo ***ebonita2_res_pd*.wav**, o sinal residual reconstruído da fala do arquivo ***ebonita2*_res_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais residuais original e o reconstruído pelo sistema WI (padrão).

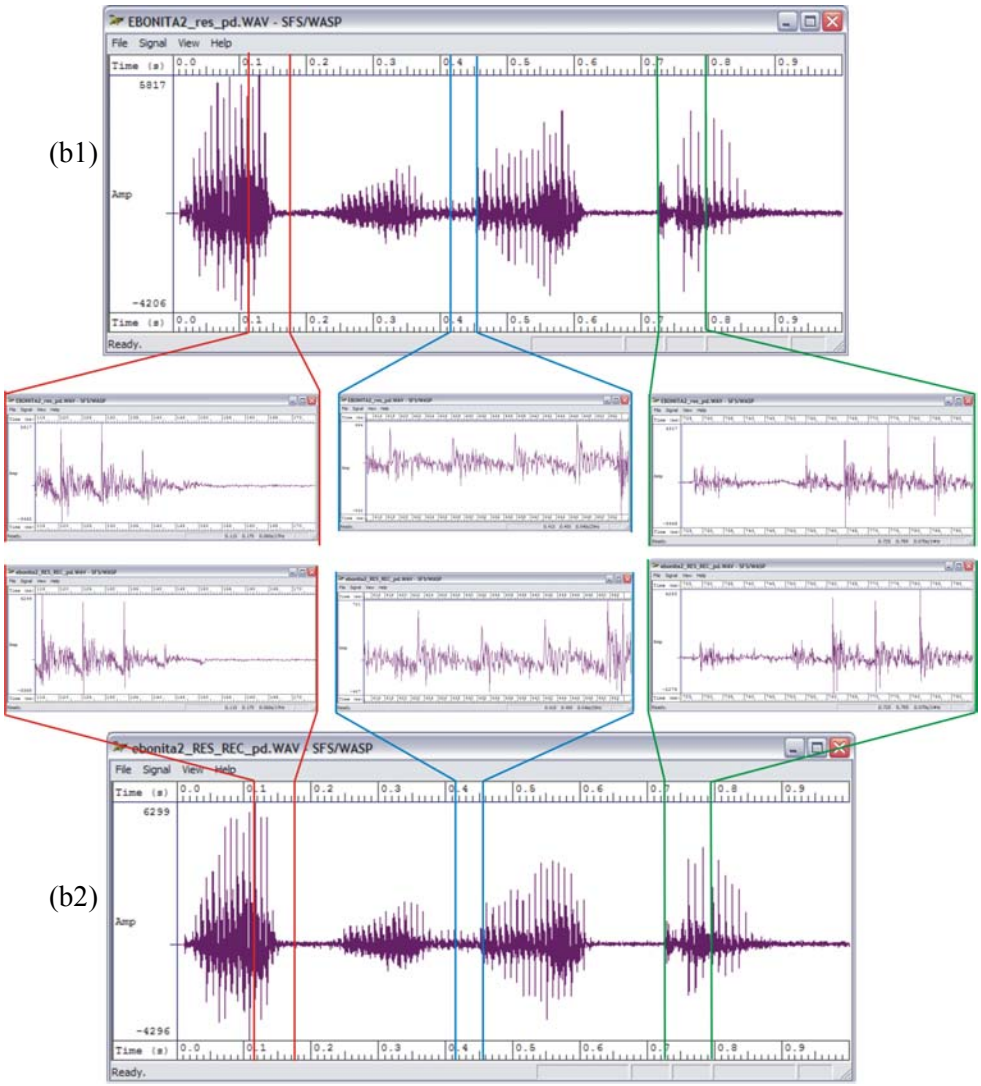

Figura 5.41 – Gráfico das formas de onda (amplitude x tempo) para os sinais residuais (locutor adulto): (b1) Sinal residual original da fala: do arquivo ***ebonita2*_res_pd.wav;** (b2) Sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo ***ebonita2*_res_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (b1) e (b2).

Na Figura 5.41, comparando os gráficos (b1) e (b2), observa-se que existem semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados, correspondentes aos segmentos sonoros da fala, também mostram grande similaridade nas formas de ondas.

5.2.4.3.3 Gráficos das formas de onda (locutora - infantil)

I - Expressão de fala (locutora - infantil): "**casa**"

(a) A Figura 5.42 mostra o sinal da fala original do arquivo ***casa3*.wav**, o sinal da fala reconstruído do arquivo ***casa3*_sin_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais original e reconstruído pelo sistema de análise – síntese WI (padrão).

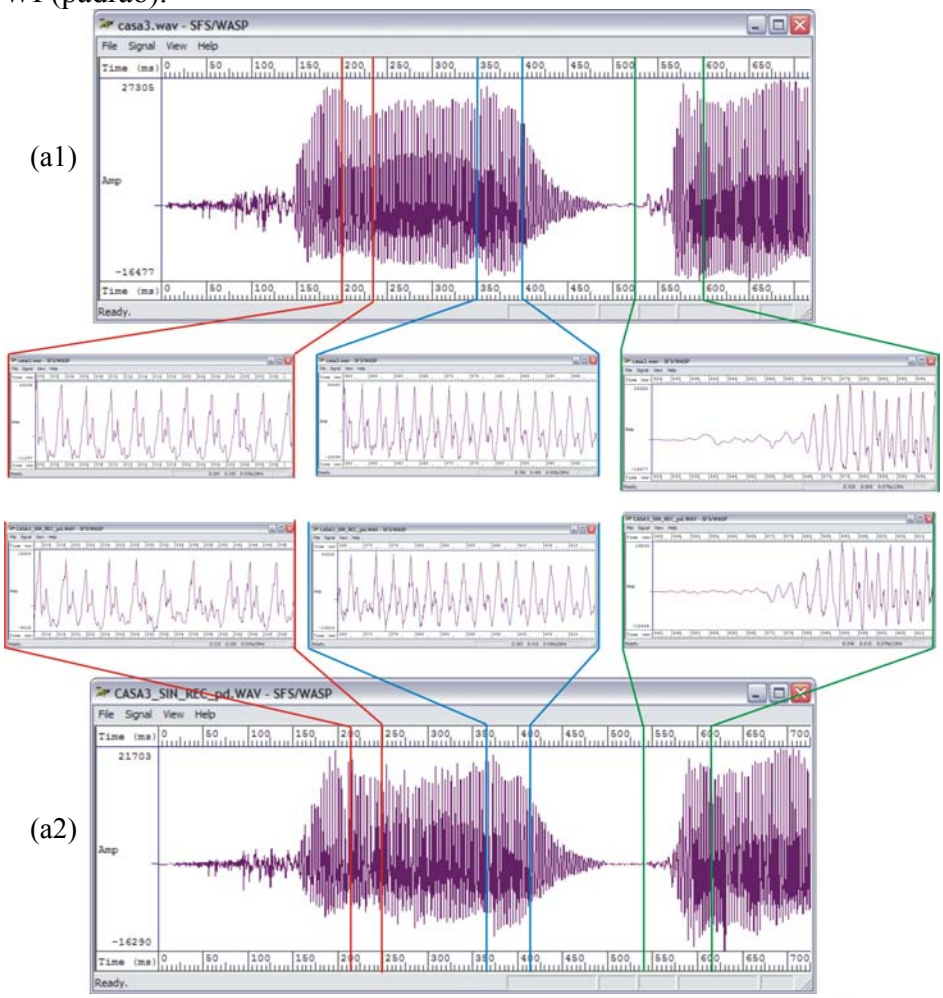

Figura 5.42 – Gráfico das formas de onda (amplitude x tempo) para os sinais (locutora infantil): (a1) Sinal da fala original: do arquivo **casa3.wav;** (a2) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo **casa3_sin_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (a1) e (a2).

Na Figura 5.42, comparando os gráficos (a1) e (a2), observa-se que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados também mostram grande similaridade através das forma de ondas principalmente nas regiões sonoras.

(b) A Figura 5.43 mostra o sinal residual original da fala do arquivo ***casa3_res_pd*.wav**, o sinal residual reconstruído da fala do arquivo ***casa3*_res_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais residuais original e o reconstruído pelo sistema WI (padrão).

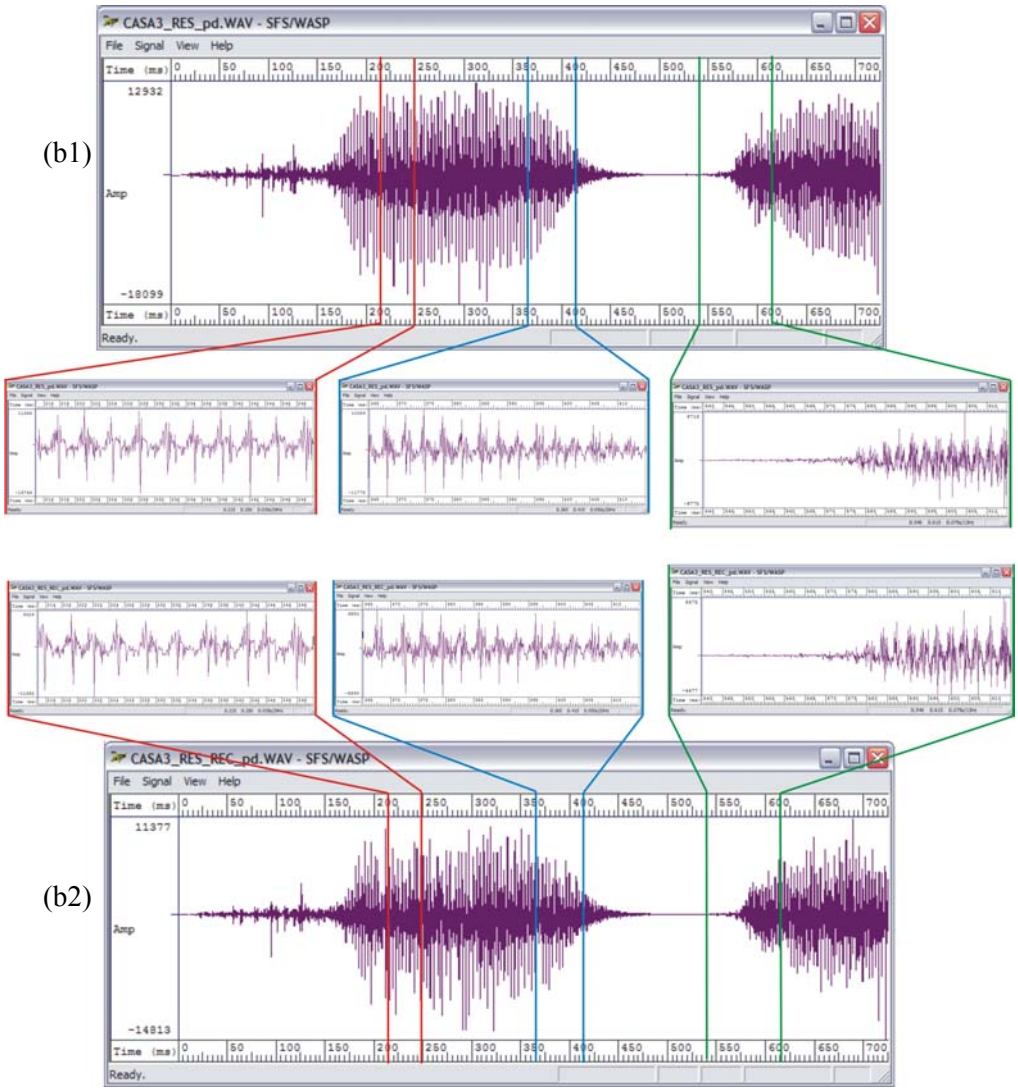

Figura 5.43 – Gráfico das formas de onda (amplitude x tempo) para os sinais residuais (locutora infantil): (b1) Sinal residual original da fala: do arquivo **casa3_res_pd.wav;** (b2) Sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo **casa3_res_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (b1) e (b2).

Na Figura 5.43, comparando os gráficos (b1) e (b2), observa-se que existem semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos, mas em (b2) o sinal se apresenta mais ruidoso. Os trechos ampliados, correspondentes aos segmentos sonoros da fala, também mostram grande similaridade nas formas de ondas.

II - Expressão de fala (locutora - infantil): "**é bonita**"

(a) A Figura 5.44 mostra o sinal da fala original do arquivo ***ebonita3*.wav**, o sinal da fala reconstruído do arquivo ***ebonita3*_sin_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais original e reconstruído pelo sistema WI (padrão).

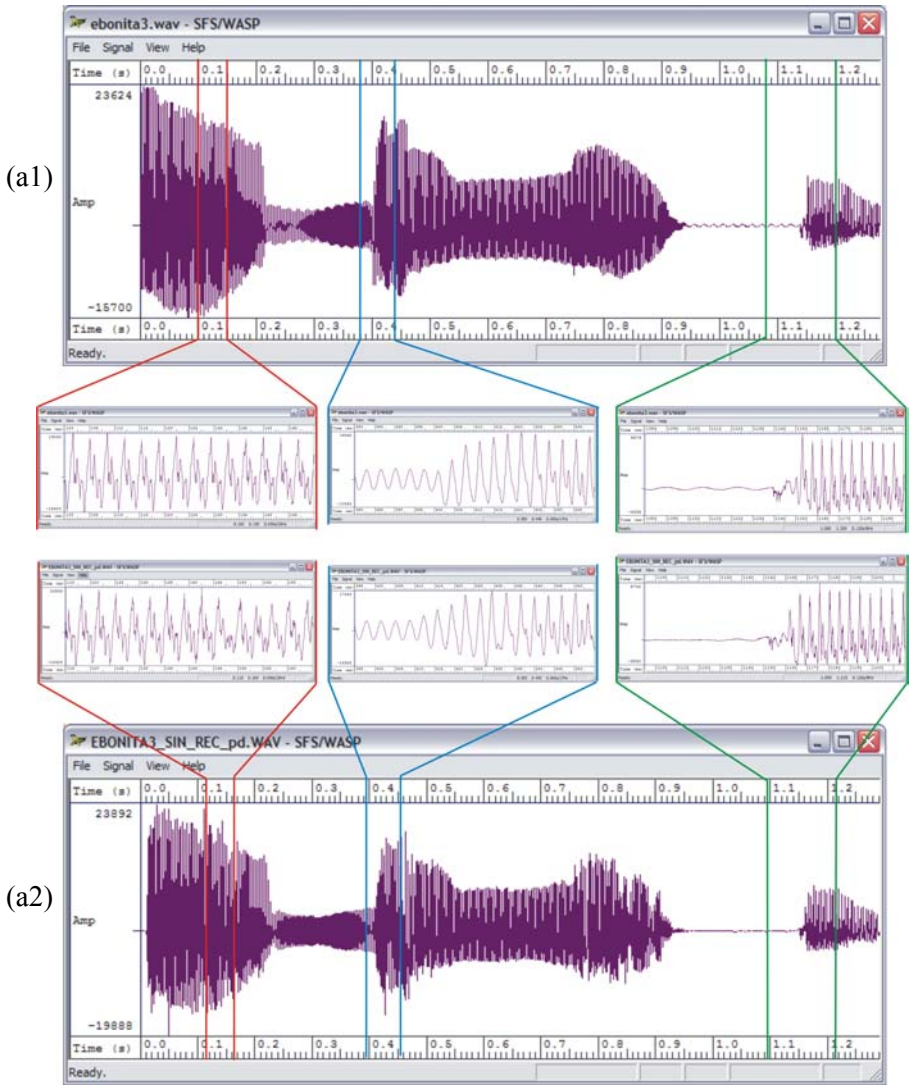

Figura 5.44 – Gráfico das formas de onda (amplitude x tempo) para os sinais (locutora infantil): (a1) Sinal da fala original: do arquivo **ebonita3.wav;** (a2) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo **ebonita3_sin_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (a1) e (a2).

Na Figura 5.44, comparando os gráficos (a1) e (a2), observa-se que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Entretanto, na região final do sinal, correspondente ao fonema /i/ nota-se diferença entre as amplitudes das formas de ondas. Os trechos ampliados também mostram grande similaridade através das formas de ondas, principalmente nas regiões sonoras.

(b) A Figura 5.45 mostra o sinal residual original da fala (locutora - infantil) do arquivo ***ebonita3_res_pd*.wav**, o sinal residual reconstruído da fala do arquivo ***ebonita3*_res_rec_pd.wav** e as ampliações de três trechos dos sinais em posições correspondentes entre os sinais residuais original e o reconstruído pelo sistema WI (padrão).

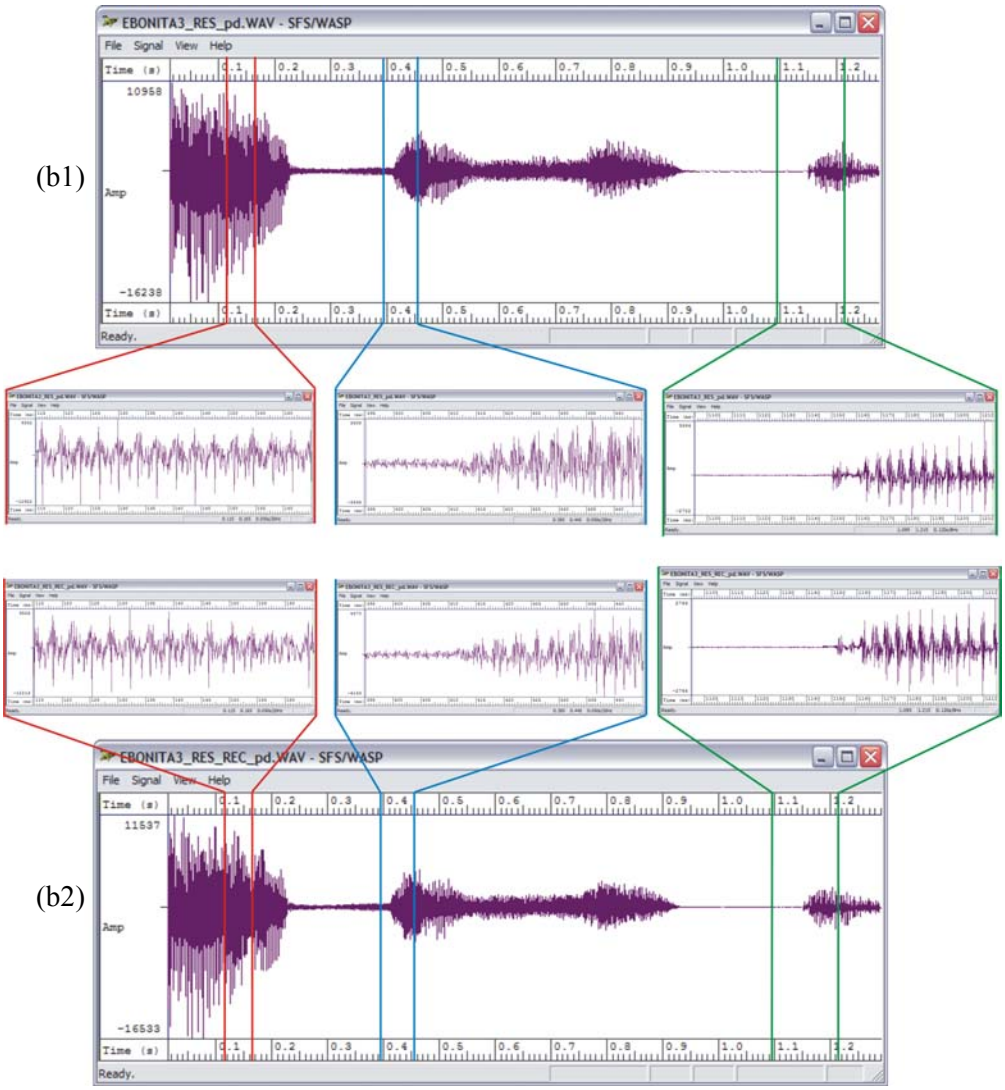

Figura 5.45 – Gráfico das formas de onda (amplitude x tempo) para os sinais residuais (locutora infantil): (b1) Sinal residual original da fala: do arquivo ***ebonita3*_res_pd.wav;** (b2) Sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): do arquivo ***ebonita3*_res_rec_pd.wav** (correspondente); e as ampliações de três trechos selecionados nas posições correspondentes indicadas nos dois sinais em (b1) e (b2).

Na Figura 5.45, comparando os gráficos (b1) e (b2), observa-se que existem semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Os trechos ampliados, correspondentes aos segmentos sonoros da fala, também mostram grande similaridade através das formas de ondas.

5.2.5 Avaliação do sistema de análise - síntese WI (com ajuste através de interpolação das CW's em posições regulares) ou sistema de análise - síntese WI (com interpolação)

Nesta seção são apresentados os objetivos do sistema de análise - síntese WI (com interpolação CW's em posições regulares) e os resultados preliminares apresentados por este sistema para uma comparação com o sistema de análise – síntese WI (padrão).

Para comparação são apresentados os resultados do processamento dos sinais dos arquivos (casa2.wav, ebonita2.wav e trechos do sinal casa2.wav) para os dois sistemas de análise – síntese WI. Para todos sinais são apresentadas as medidas PESQ_MOS e SNRSEG – NCCF; e para o sinal do arquivo casa2.wav e também para os trechos desse mesmo arquivo são apresentados os resultados da aplicação do algoritmo SNRSEG-NCCF na forma de gráficos, mostrando o deslocamento (ou defasagem), a correlação cruzada normalizada e a relação sinal ruído por segmento. Também são apresentados os gráficos das formas de onda dos sinais originais e dos sinais reconstruídos para o sistema de análise - síntese WI (padrão) e para o sistema de análise - síntese WI (com interpolação).

5.2.5.1 Os propósitos do sistema de análise - síntese WI (com interpolação das CW's em posições regulares)

O objetivo principal do sistema é tentar resolver o problema da localização das "Characteristic Waveforms" - (CW's) em posições deslocadas das posições regulares. Para isto são levantadas algumas considerações sobre o processo de localização e extração das *formas de onda características*, CW's, e sobre a estimação do pitch na técnica de interpolação das ondas, técnica WI:

- Em geral, na técnica WI, o processo de estimação do pitch é executado uma vez por quadro e interpolado linearmente por sub quadros, resultando nos valores de pitch utilizados para localizar as CW's nas posições regulares. No sinal residual as formas de ondas dos ciclos de pitch, as CW's, geralmente variam no formato e no valor do período de pitch de um ciclo para o outro. Este processo, portanto, não resulta em um

valor preciso para o pitch por quadro, e menos ainda para os valores interpolados por sub quadro;

- Em geral, a forma de onda do ciclo do pitch é localizada em uma posição deslocada da posição regular própria, onde a CW deveria ser localizada e extraída. Mas a CW é considerada como se fosse localizada corretamente na posição regular. Os parâmetros relativos a CW são então transmitidos ao decodificador onde novamente a CW é considerada na posição regular durante o processo de interpolação;
- Nos trabalhos de E. Choy [14] e de W. B. Kleijn [15] é permitido uma certa flexibilidade durante a localização da posição das CW's em torno da posição regular própria onde a CW deveria ser localizada. A posição pode variar em uma faixa de $[-\varepsilon_{máx}, +\varepsilon_{máx}]$ em torno da posição regular. Com estas considerações Choy [14] relata que experimentos informais confirmaram que $\varepsilon_{máx}$ pode assumir um valor de até 16 amostras sem afetar a qualidade perceptual da fala reconstruída.
- A técnica WI durante o processo de reconstrução do sinal, na decodificação, requer a cada instante no tempo (espaçados por um período de amostragem) uma CW com a respectiva fase para determinar a correspondente amostra do sinal residual reconstruído. Para atender esta exigência, no decodificador é realizada a interpolação das CW's em cada instante de tempo entre as CW's recebidas em posições regulares. Desta forma, se a CW foi localizada em uma posição deslocada da posição regular, mas foi considerada em tal posição e transmitida ao decodificador, então em uma visão mais apurada do processo, ocorrerá alguma distorção na forma de onda durante a reconstrução do sinal residual.

5.2.5.2 Os processos para a melhoria na localização e na extração das CW's

Partindo das considerações anteriores e na tentativa de diminuir as distorções causadas pelas CW's localizadas em posições deslocadas das posições regulares, foram visualizados três procedimentos que incorporados à técnica WI podem resultar na melhoria na reconstrução do sinal, durante o estágio de síntese (ou decodificação). Assim, os procedimentos são:

- aprimorar o processo de localização e extração das formas de onda características, as CW's;
- a partir da localização mais precisa das CW's, ajustar a estimação do pitch;
- e a partir das CW's localizadas com maior precisão e do pitch estimado ajustado, obter as CW's nas posições regulares através de interpolação.

Através do aprimoramento dos processos, *localização das CW's*, *ajuste do pitch estimado* e *determinação das CW's em posição regular*, obtêm-se uma análise mais apurada e uma

codificação mais eficiente que resultam em uma melhor reconstrução do sinal residual e síntese do sinal da fala na técnica WI.

Observação: N*o Apêndice A desta tese é apresentada uma proposta detalhada com a aplicação dos processos para melhoria do sistema de análise – síntese (padrão).*

5.2.5.3 Os processos realizados para a melhoria na localização e na extração das CW's no sistema de análise – síntese WI (com ajuste através de interpolação das CW's em posições regulares)

Com a intenção de resumir a pesquisa por causa da limitação de tempo para esta tese, mas com o objetivo de apresentar algum resultado que comprove estas hipóteses, ou parte delas, resolveu-se simplificar, excluindo o processo do *ajuste do pitch estimado*. Desta forma o sistema de análise – síntese WI (com ajuste através de interpolação das CW's em posições regulares) foi implementado considerando que:

a. As CW's são localizadas conforme o sistema de análise – síntese WI (padrão);
b. A partir das CW's localizadas, próximas as posições regulares, determina-se as CW's nas posições regulares, através de interpolação no domínio de Fourier;
c. Os parâmetros relativos às CW's interpoladas, em posições regulares, são obtidos e transmitidos ao sistema de síntese da mesma forma que foi realizado no sistema de análise – síntese WI (padrão);
d. O sistema com as modificações consideradas, denominado de *sistema de análise – síntese WI com interpolação das CW's em posições regulares,* será simplificado para *sistema de análise – síntese WI (com interpolação);*
e. As interpolações são realizadas, no domínio de Fourier, da mesma forma com foi descrito no processo D233 no capítulo 4 desta tese;
f. Todos os itens anteriores são descritos em detalhes no apêndice A desta tese.

5.2.5.4 Os resultados: medida PESQ_MOS e SNRSEG – NCCF para o sistema de análise – síntese WI (com interpolação)

Nesta seção são apresentados os resultados PESQ_MOS e da SNRSEG-NCCF relativos ao processamento dos sinais da fala nos sistemas de análise – síntese WI (padrão) e (com interpolação). O objetivo é avaliar a versão preliminar do sistema de análise – síntese WI (com interpolação) pelos resultados iniciais, comparando os sinais da fala reconstruídos pelos dois sistemas para um mesmo sinal da fala original, utilizando as medidas PESQ_MOS e SNRSEG-NCCF.

Os resultados com as medidas PESQ_MOS e SNRSEG - NCCF são mostrados em tabelas e gráficos onde o nome dos arquivos com os sinais processados pelo *sistema de análise – síntese WI (padrão*) têm a indicação abreviada *"pd"* e pelo *sistema de análise – síntese WI (com interpolação)* a indicação *"c_interp",* conforme a notação adotada e mostrada na Figura 5.3. A Figura 5.46 mostra o esquema utilizado para comparar os resultados PESQ_MOS e SNRSEG-NCCF obtidos com as simulações do *sistema de análise – síntese WI (com interpolação)* e do *sistema de análise – síntese WI (padrão)*. Assim, para os nomes dos arquivos com os sinais são considerados:

(a) Para o sinal da fala:

- *Padrão de referência*:

  Sinal original da fala => ***nomearq*.wav**;

  Trecho do sinal original da fala => ***nomearq_tc1*.wav**; ***nomearq_tc2*.wav**; ***nomearq_tc3*.wav.**

- *Sinal em teste (degradado)*:

  Sinal da fala reconstruído (sintetizado) => ***nomearq*_sin_rec_pd.wav** ou ***nomearq*_sin_rec_ c_interp.wav**;

  Trecho do sinal da fala reconstruído (sintetizado) =>

  ***nomearq_sin_rec_pd_tc1*.wav** ou ***nomearq*_sin_rec_c_interp *tc1*.wav**;

  ***nomearq_sin_rec_pd_ tc2*.wav** ou ***nomearq*_sin_rec_c_interp *tc2*.wav**;

  ***nomearq_sin_rec_pd_ tc3*.wav** ou ***nomearq*_sin_rec_ c_interp *tc3*.wav**;

(b) Para os sinais residuais:

- *Padrão de referência*: Sinal residual da fala (lado da análise) => ***nomearq*_res_pd.wav** ou ***nomearq*_res_c_interp.wav;**
- *Sinal em teste (degradado)*: Sinal residual reconstruído (lado da síntese) => ***nomearq*_res_rec_pd.wav** ou ***nomearq*_res_rec_ c_interp.wav**.

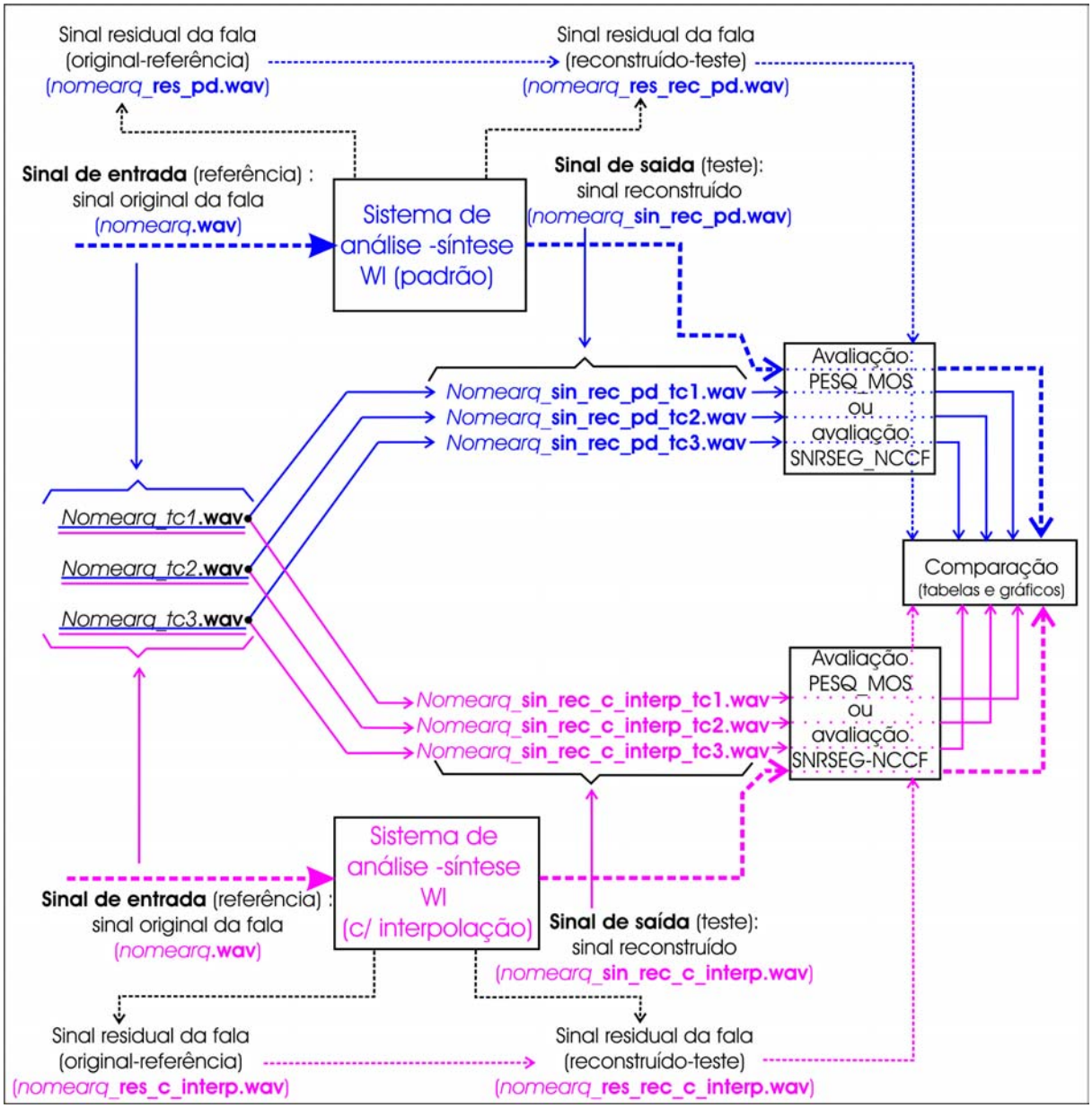

Figura 5.46 - Diagrama esquemático para a comparação das simulações do sistema de análise – síntese WI (com interpolação) com o sistema de análise – síntese WI (padrão) pelos resultados PESQ_MOS e SNRSEG-NCCF.

Assim, foram processados nos sistemas, os sinais originais da fala nos arquivos casa2.wav e ebonita2.wav obtendo-se os sinais reconstruídos: nos arquivos casa2_sin_rec_pd.wav e ebonita2_sin_rec_pd.wav para o sistema (padrão); e nos arquivos casa2_sin_rec_c_interp.wav e ebonita2_sin_rec_c_interp.wav para o sistema (com interpolação). Aos sinais originais e reconstruídos correspondentes aplica-se o algoritmo PESQ e o algoritmo da SNRSEG-NCCF obtendo-se as medidas PESQ_MOS e SNRSEG-NCCF, respectivamente. Os mesmos procedimentos foram realizados também com os sinais residuais. Os resultados são apresentados na Tabela 5.8.

Para uma avaliação mais detalhada, conforme mostra o esquema na Figura 5.46, foram selecionados trechos do sinal da fala no arquivo casa2.wav e gravados nos arquivos casa2_tc1.wav, casa2_tc2.wav e casa2_tc3.wav. Nos sinais reconstruídos nos arquivos casa2_sin_rec_pd.wav e casa2_sin_rec_c_interp.wav foram localizados os trechos correspondentes aos trechos do sinal original casa2.wav. Aos sinais originais e reconstruídos dos trechos de sinais aplica-se também o algoritmo PESQ e o algoritmo SNRSEG-NCCF, obtendo-se também as medidas PESQ_MOS e da SNRSEG-NCCF respectivamente, que são apresentadas na Tabela 5.9. Também são apresentados os resultados com a aplicação do algoritmo SNRSEG-NCCF, mostrando na forma de gráficos: a defasagem, a correlação cruzada normalizada e a relação sinal ruído por segmento.

A seguir são apresentados os resultados PESQ_MOS e SNRSEG-NCCF relativos aos processamentos nos sistemas de análise – síntese WI, padrão e com interpolação, para os arquivos com os sinais da fala, para o locutor – adulto, armazenados nos arquivos *casa2*.wav e ebonita2.wav. Os resultados são apresentados na Tabela 5.8.

Tabela 5.8 – Resultados PESQ_MOS e SNRSEG-NCCF. Locutor – adulto- Sistemas de análise/síntese WI (padrão) e (com interpolação).

Locutor – adulto					
Expressão da fala	**Sinal original (referência)**	**Sinal reconstruído (em teste)** *Sistema de análise – síntese WI: padrão* => pd**;** *c/interpolação* => c_interp**.**	**PESQ_MOS (–0,5 a 4,5)**	**SNRSEG (NCCF)** (dB)	**d'med / d'med_dif_abs** (amostras)
"casa"	casa2.wav	*casa2*_sin_rec_pd.wav	**3,916**	**10,619228**	**-26 / 4**
		*casa2*_sin_rec_interp.wav	**3,397**	**11,340422**	**-6 / 10**
	*casa2*_res.wav	*casa2*_res_rec_pd.wav	**3,780**	**2,978854**	**22 / 12**
		*casa2*_res_rec_ interp.wav	**3,348**	**2,246722**	**-9 / 11**
"é bonita"	*ebonita2*.wav	*ebonita2*_sin_rec_pd.wav	**3,375**	**9,363902**	**-19 / 11**
		*ebonita2*_sin_rec_interp.wav	**2,952**	**9,723617**	**-3 / 13**
	*ebonita2*_res.wav	*ebonita2*_res_rec_pd.wav	**3,543**	**2,111374**	**8 / 21**
		*ebonita2*_res_rec_interp.wav	**3,273**	**0,687373**	**-2 / 17**

Os resultados mostrados na Tabela 5.8 indicam que:

I - Em relação à medida **PESQ_MOS**:

(a) Para os sinais da fala e sinais residuais da fala: os resultados PESQ_MOS são melhores para o sistema de análise –síntese WI (padrão);

(b) Para a expressão "casa": os resultados PESQ_MOS para os sinais da fala são maiores do que os resultados PESQ_MOS para os sinais residuais da fala;

(c) Para a expressão "é bonita": os resultados PESQ_MOS para os sinais da fala são menores do que os resultados PESQ_MOS para os sinais residuais da fala;

II - Em relação à medida **SNRSEG-NCCF**:

(a) Para os sinais da fala: esta medida apresentou maiores valores no sistema de análise – síntese WI (com interpolação); Estes resultados indicam que os sinais reconstruídos da fala no sistema de análise – síntese WI (com interpolação) apresentam uma melhor reconstrução por segmento que o sistema de análise –síntese WI (padrão).

(b) Para os sinais residuais da fala: esta medida foi maior no sistema de análise – síntese WI (padrão);

III - Em relação à medida **d'med_dif_abs**:

(a) Para os sinais da fala as medidas d'med_dif_abs indicam que o sistema de análise – síntese WI (com interpolação) apresenta uma defasagem maior do que no sistema de análise – síntese (padrão);

(b) Para os sinais residuais da fala: as medidas d'med_dif_abs indicam que o sistema de análise – síntese WI (com interpolação) apresenta uma defasagem menor do que no sistema de análise – síntese (padrão);

Os resultados PESQ_MOS mostram que a qualidade perceptual para os sinais da fala é melhor no sistema de análise – síntese (padrão).

Os resultados SNRSEG-NCCF mostram que os sinais da fala no sistema de análise – síntese WI (com interpolação) têm uma melhor reconstrução por segmento.

Os resultados das medidas d'med_dif_abs indicam que para os sinais da fala o sistema de análise – síntese WI (com interpolação) apresenta um grau de defasagem maior do que no sistema de análise – síntese (padrão).

Estas conclusões sinalizam que o sistema de análise – síntese WI (com interpolação) poderá ter melhores resultados que o sistema de análise –síntese WI (padrão) se houver uma correção no defasamento do sinal.

A seguir são apresentados os resultados PESQ_MOS e da SNRSEG-NCCF relativos ao processamento do sinal completo do arquivo casa2.wav e dos trechos selecionados no arquivo casa2.wav (gravados nos arquivos casa2_tc1.wav, casa2_tc2.wav e casa2_tc3.wav) nos sistemas de análise – síntese WI (padrão) e (com interpolação). Os resultados são apresentados na Tabela 5.9.

Tabela 5.9 - Locutor – adulto- Sistema análise - síntese WI - Trechos selecionados do sinal no arquivo casa2.wav.

Locutor – adulto – Expressão da fala "casa"					
	Sinal original (referência)	**Sinal reconstruído (em teste)** *Sistema de análise – síntese WI:* *padrão* => pd**;** *c/interpolação* => c_interp**.**	**PESQ_MOS** (–0,5 a 4,5)	**SNRSEG (NCCF)** (dB)	**d'med / d'med_dif_abs** (amostras)
Sinal completo	casa2.wav	*casa2*_sin_rec_pd.wav	**3,916**	**10,619228**	**-26 / 4**
		*casa2*_sin_rec_c_interp.wav	**3,397**	**11,340422**	**-6 / 10**
Trecho 1	casa2_tc1.wav	*casa2*_sin_rec_pd_tc1.wav	**2,182**	**12,058152**	**-1 / 0**
		*casa2*_sin_rec_c_interp_tc1.wav	**3,860**	**14,030409**	**13 / 4**
Trecho 2	casa2_tc2.wav	*casa2*_sin_rec_pd_tc2.wav	**3,620**	**11,516921**	**-29 / 3**
		*casa2*_sin_rec_c_interp_tc2.wav	**3,505**	**12,994504**	**-8 / 5**
Trecho 3	casa2_tc3.wav	*casa2*_sin_rec_pd_tc3.wav	**4,011**	**12,511552**	**0 / 0**
		*casa2*_sin_rec_c_interp_tc3.wav	**4,120**	**9,492401**	**10 / 7**

Os resultados mostrados na Tabela 5.9 indicam que:

I - Em relação à medida **PESQ_MOS (sinais da fala)**:

(a) Para o trecho 2 e para o arquivo completo "casa2.wav" com sinais da fala: os resultados PESQ_MOS são melhores para o sistema de análise – síntese WI (padrão);

(b) Para os trechos 1 e 3 com sinais da fala: os resultados PESQ_MOS são melhores para o sistema de análise –síntese WI (com interpolação);

II - Em relação à medida **SNRSEG-NCCF**:

Para os trechos 1 e 2 e o sinal completo "casa": o sistema de análise – síntese WI (com interpolação) apresentou melhores resultados em relação à medida **SNRSEG-NCCF**, enquanto que para o trecho 3 o sistema de análise – síntese WI (padrão) apresentou melhores resultados; estes resultados indicam que os sinais reconstruídos da fala no sistema de análise – síntese WI (com interpolação) podem apresentar uma melhor reconstrução por segmento que o sistema de análise – síntese WI (padrão).

III - Em relação à medida **d'med_dif_abs**:

O sistema de análise – síntese WI (padrão) apresentou melhores resultados em relação as medidas d'med_dif_abs (valores menores) para todos os trechos e para o arquivo com os sinais da fala "casa"; estes resultados indicam que esses sinais reconstruídos da fala no sistema de análise – síntese WI (padrão) apresentam um menor grau de defasagem que o sistema de análise – síntese WI (com interpolação).

Os resultados PESQ_MOS mostram que a qualidade perceptual para os sinais da fala é melhor no sistema de análise – síntese (padrão) para o sinal completo e para o trecho 2

enquanto que para os trechos 1 e 3 é melhor para o sistema de análise – síntese WI (com interpolação).

Os resultados SNRSEG-NCCF mostram que os sinais da fala no sistema de análise – síntese WI (com interpolação) têm uma melhor reconstrução por segmento para os trechos 1 e 2 e para o sinal completo "casa".

Os resultados das medidas d'med_dif_abs indicam que para os sinais da fala o sistema de análise – síntese WI (com interpolação) apresenta um grau de defasagem maior do que no sistema de análise – síntese (padrão).

Estas conclusões sinalizam que o sistema de análise – síntese WI (com interpolação) poderá ter melhores resultados que o sistema de análise –síntese WI (padrão) se houver uma correção no defasamento do sinal.

De forma geral as Tabelas 5.8 e 5.9 apresentam resultados que permite concluir que o sistema de análise – síntese WI (com interpolação) apresenta uma reconstrução do sinal da fala por segmento melhor do que o sistema de análise –síntese WI (padrão). Por outro lado o sistema de análise – síntese WI (com interpolação) poderá ter melhores resultados que o sistema de análise –síntese WI (padrão) se houver alguma alteração no modelo para correção no defasamento do sinal.

Na seqüência são apresentados os resultados da aplicação do algoritmo SNRSEG-NCCF, relativos aos mesmos processamentos, através dos gráficos para o deslocamento, a correlação cruzada normalizada e a relação sinal ruído por segmento nas Figuras 5.47, 5.48, 5.49, 5.50, 5.51, 5.52, 5.53, 5.54, 5.55, 5.56, 5.57 e 5.58.

A Figura 5.59 mostra a localização dos trechos selecionados nos arquivos casa2.wav, casa2_sin_rec_pd.wav e casa2_sin_rec_c_interp.wav. Os trechos com os sinais da fala foram extraídos em posições correspondentes nos três arquivos. Por uma questão didática a Figura 5.59 foi repetida na Figura 5.47.

A Figura 5.47 mostra os deslocamentos d' (em amostras), obtidos para cada segmento do sinal da fala no arquivo em teste, casa2_sin_rec_pd.wav ou casa2_sin_rec_c_interp.wav em relação ao sinal original casa2.wav, com a aplicação do algoritmo da SNRSEG-NCCF para segmentos com M = 99 amostras.

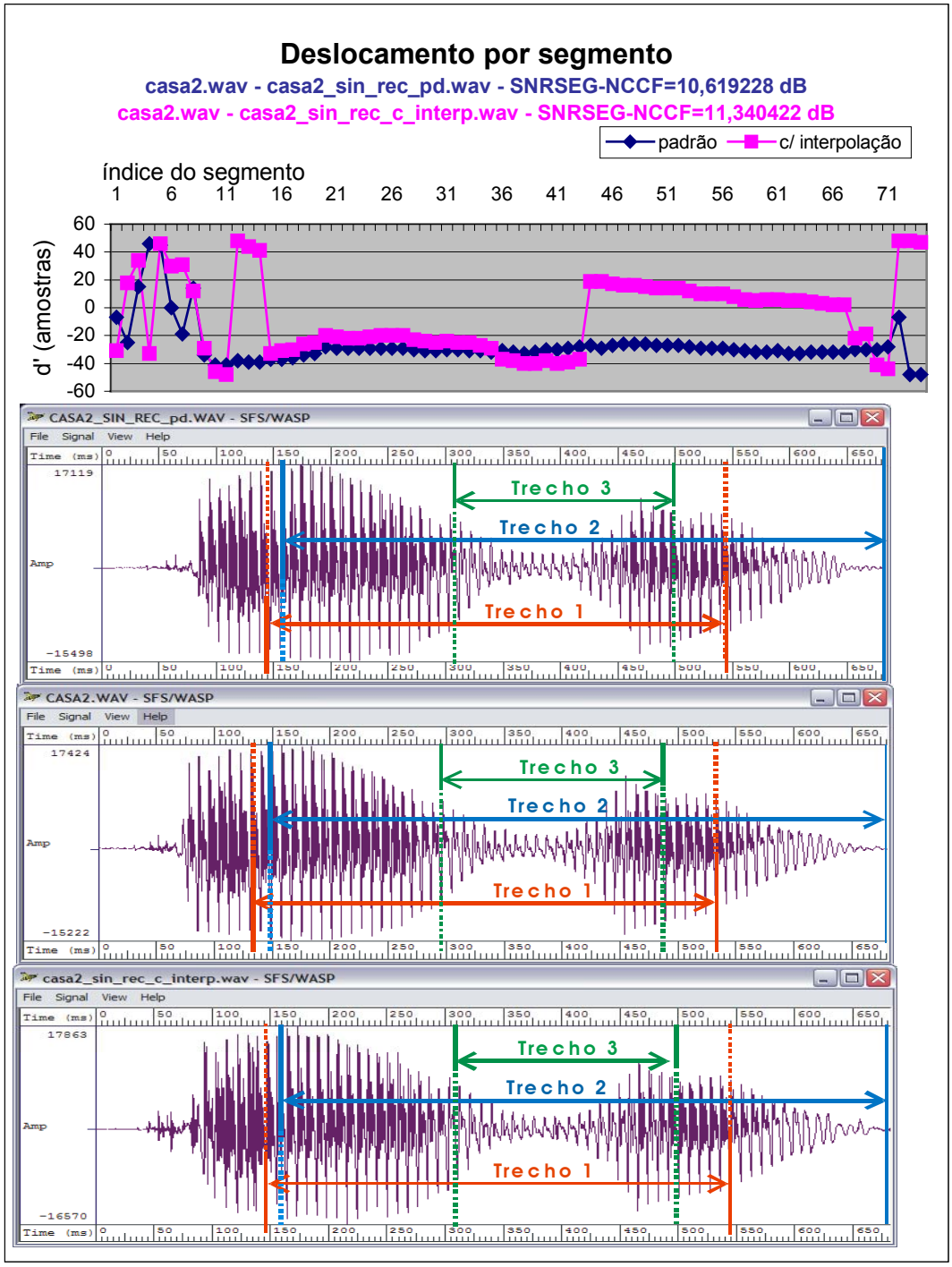

Figura 5.47 – Deslocamentos em número de amostras, d', para os segmentos do sinal reconstruído, *casa2*_sin_rec_pd.wav e casa2_sin_rec_c_interp.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original, *casa2*.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra d' para o sistema de analise – síntese WI (padrão) e na cor magenta mostra d' para o sistema de análise – síntese WI (com interpolação). Para a visualização da posição relativa dos segmentos são mostrados também os gráficos das formas de onda amplitude x tempo para os sinais da fala nos arquivos envolvidos.

Na Figura 5.47, no primeiro gráfico de cima para baixo, observa-se que o sistema de análise – síntese WI (padrão), na cor azul, apresenta uma seqüência de deslocamentos por segmento mais uniforme que é indicado pelo valor d'med_dif_abs = 4 amostras (conforme a Tabela 5.8). Para o sistema de análise – síntese WI (com interpolação) observa-se dois trechos onde a seqüência é uniforme, mas com uma descontinuidade entre eles. O primeiro trecho segue de forma aproximada os valores para os deslocamentos do sistema padrão, mas o segundo trecho se afasta da seqüência uniforme em azul. Observa-se então que os sinais processados no sistema de análise – síntese WI (com interpolação) para o sinal completo "casa" apresentam-se mais defasados que no sistema de análise – síntese WI (padrão).

A Figura 5.48 mostra as correlações cruzadas normalizadas, R (NCCF), entre os segmentos no arquivo *casa2*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd.wav ou no arquivo casa2_sin_rec_c_interp.wav calculadas com a aplicação do algoritmo da SNRSEG-NCCF para segmentos com M = 99 amostras.

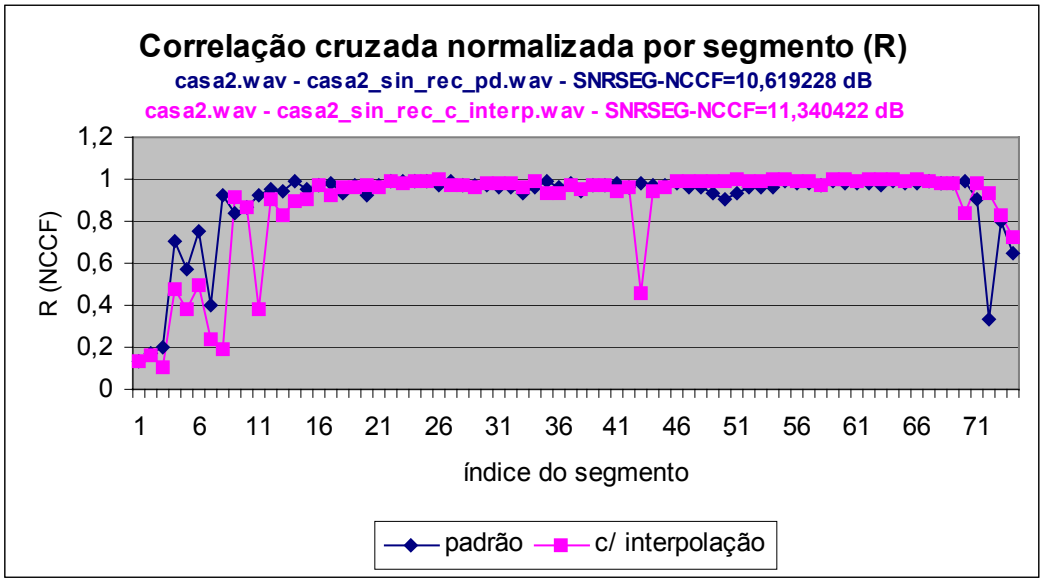

Figura 5.48 – Correlação cruzada normalizada, R (NCCF), entre os segmentos no arquivo *casa2*.wav e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd.wav ou no arquivo casa2_sin_rec_c_interp.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra R para o sistema de análise – síntese WI (padrão) e na cor magenta mostra R para o sistema de análise – síntese WI (com interpolação).

Na Figura 5.48 observa-se nos trechos mais uniformes para a correlação R resultados melhores e maiores, por uma pequena margem, para o sistema de análise – síntese WI (com interpolação) indicados também pela medida SNRSEG-NCCF = 11,34042 dB comparada com 10,619228 dB para o sistema de analise – síntese WI (padrão). Essa figura mostra

também que para o segmento número 43 existe uma descontinuidade que corresponde à mesma posição para a descontinuidade dos deslocamentos indicados na Figura 5.47.

A Figura 5.49 mostra as relações sinal ruído (SNR), entre os segmentos no arquivo *casa2*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd.wav ou no arquivo casa2_sin_rec_c_interp.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para segmentos com M = 99 amostras.

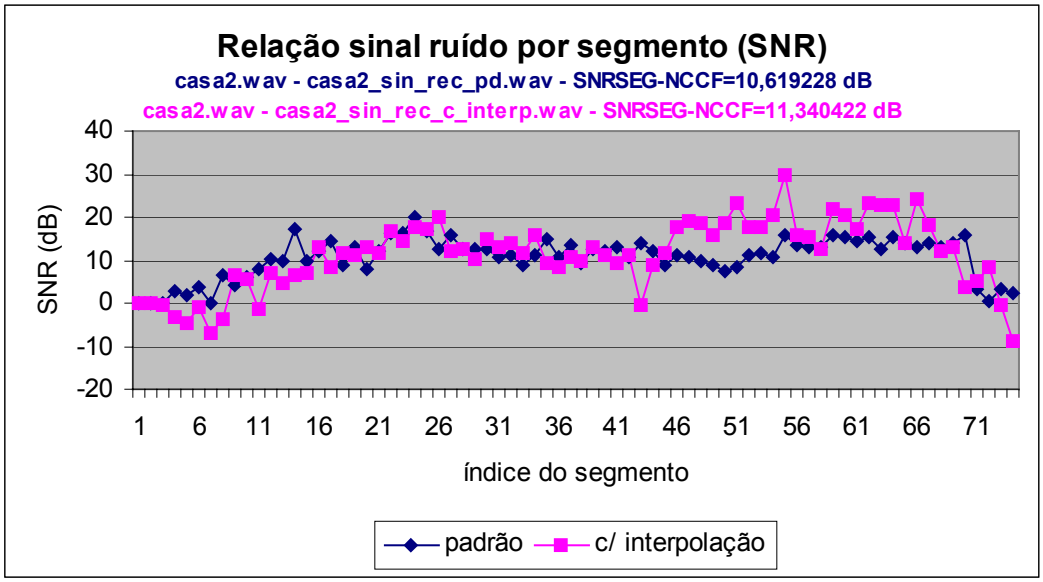

Figura 5.49 – Relação sinal ruído, SNR (dB), entre os segmento no arquivo *casa2*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd.wav ou no arquivo casa2_sin_rec_c_interp.wav. Esses resultados foram obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra SNR (dB) para o sistema de análise – síntese WI (padrão) e na cor magenta mostra SNR (dB) para o sistema de análise – síntese WI (com interpolação).

Na Figura 5.49 observa-se uma maior SNR para as regiões do sinal sonoro com maiores energias. A SNR para o sistema de análise – síntese (com interpolação) apresenta maiores resultados mostrados pela curva em magenta, o que é indicado também pelo valor SNRSEG-NCCF = 11,340422 dB comparado com 10,619228 dB para o sistema WI (padrão). As regiões que apresentam menores correlações R também apresentam SNR's menores. Estes resultados indicam que o sistema de análise – síntese WI (com interpolação) apresenta uma melhor reconstrução do sinal por segmento do que o sistema de análise – síntese WI (padrão) para o sinal do trecho 1.

A seguir são apresentados os resultados da aplicação do algoritmo SNRSEG-NCCF para a expressão da fala "casa" locutor – adulto utilizando os trechos correspondentes da fala extraídos dos arquivos *casa2*.wav (sinal original-referência), casa2_sin_rec_pd.wav e

casa2_sin_rec_c_interp.wav (sinal reconstruído-em teste). Os resultados são apresentados nos gráficos nas Figuras 5.50, 5.51, 5.52, 5.53, 5.54, 5.55, 5.56, 5.57 e 5.58.

A Figura 5.50 mostra os deslocamentos d' (em número de amostras) que foram obtidos para cada segmento do arquivo em teste casa2_sin_rec_pd_tc1.wav ou casa2_sin_rec_c_interp_tc1.wav em relação ao arquivo original casa2_tc1.wav com a aplicação do algoritmo da SNRSEG-NCCF, para segmentos com M = 99 amostras.

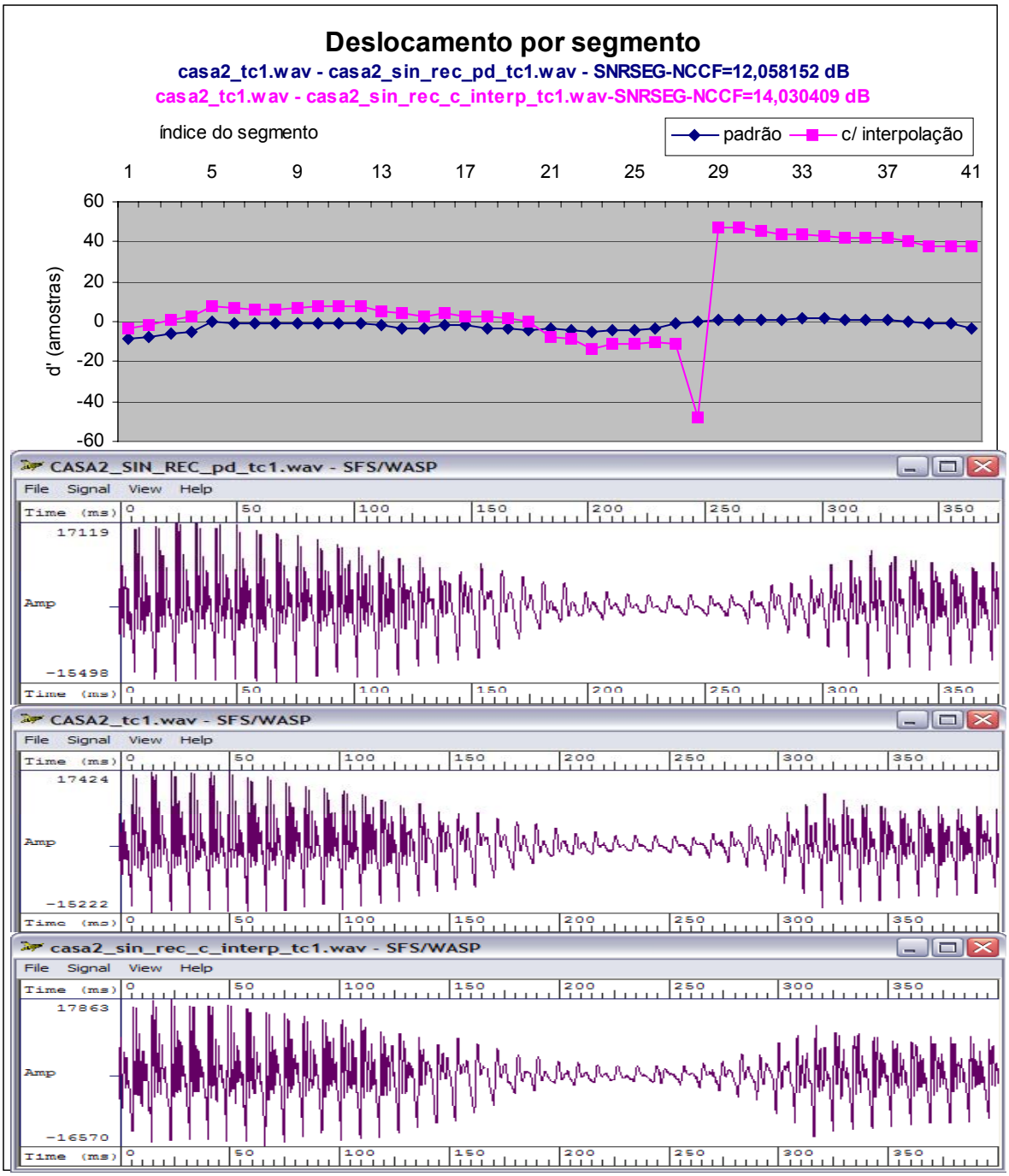

Figura 5.50 – Deslocamentos em número de amostras, d', para os segmentos do sinal reconstruído, no arquivo *casa2*_sin_rec_pd_tc1.wav ou no arquivo casa2_sin_rec_c_interp_tc1.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original, no arquivo *casa2_tc1*.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra d' para o sistema de análise – síntese WI (padrão) e na cor magenta mostra d' para o sistema de análise – síntese WI (com interpolação). Para a visualização da posição relativa dos segmentos são mostrados também os gráficos das formas de onda amplitude x tempo para os sinais da fala nos arquivos envolvidos.

Na Figura 5.50, no primeiro gráfico de cima para baixo, observa-se que o sistema de análise – síntese WI (padrão), na cor azul, apresenta uma seqüência de deslocamentos por segmento mais uniforme. Isto é indicado pelo valor d'med_dif_abs igual a 0 amostras enquanto que d'med_dif_abs é de 4 amostras para o o sistema de análise – síntese WI (com interpolação), conforme a Tabela 5.9. Para o sistema de análise – síntese WI (com interpolação) observa-se dois trechos onde a seqüência é uniforme mas com uma descontinuidade entre eles. O primeiro trecho segue de forma aproximada os valores para os deslocamentos do sistema padrão, mas o segundo trecho se afasta da seqüência uniforme em azul. Observa-se então que para os sinais processados no sistema de análise – síntese WI (com interpolação) para o trecho 1 apresentam-se mais defasados que no sistema de análise – síntese WI (padrão).

A Figura 5.51 mostra as correlações cruzadas normalizadas, R (NCCF), entre os segmentos no arquivo *casa2_tc1*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd_tc1.wav ou no arquivo casa2_sin_rec_c_interp_tc1.wav calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 99 amostras.

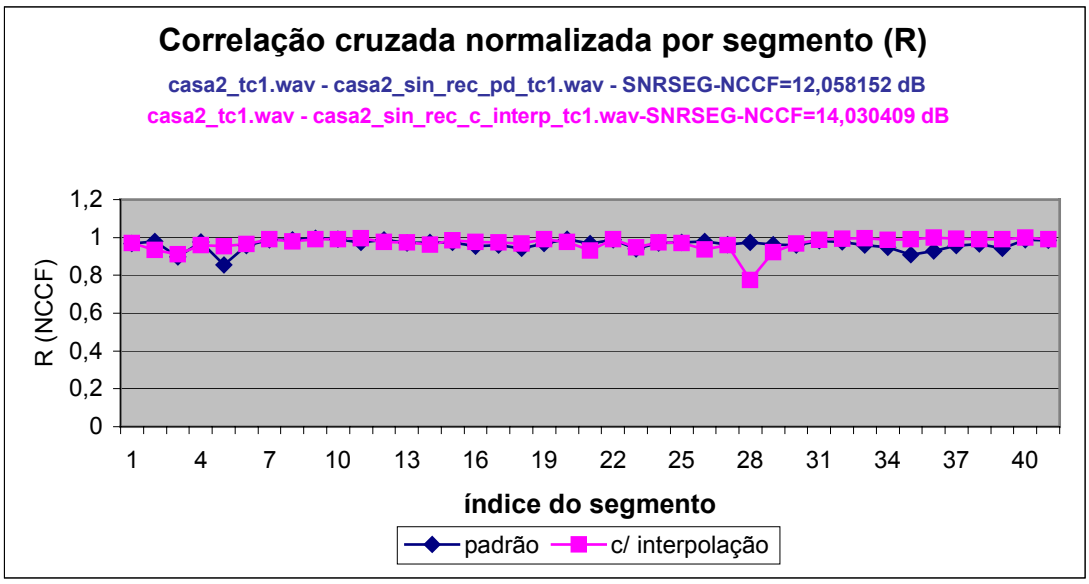

Figura 5.51 – Correlação cruzada normalizada, R (NCCF), entre os segmentos no arquivo *casa2_tc1*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd_tc1.wav ou no arquivo casa2_sin_rec_c_interp_tc1.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra R para o sistema de análise – síntese WI (padrão) e na cor magenta mostra R para o sistema de análise – síntese WI (com interpolação).

Na Figura 5.51 observa-se uma seqüência de valores mais uniformes para a correlação R com resultados maiores, por uma pequena margem, para o sistema de análise – síntese WI (com

interpolação) indicados também pela medida SNRSEG-NCCF = 14,030409 dB comparada com 12,058152 dB para o sistema de análise – síntese WI (padrão). Essa figura mostra também que para o segmento número 28 existe uma descontinuidade que corresponde à mesma posição para a descontinuidade dos deslocamentos indicados na Figura 5.50.

A Figura 5.52 mostra as relações sinal ruído, SNR (dB), entre os segmentos no arquivo *casa2_tc1*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd_tc1.wav ou no arquivo casa2_sin_rec_c_interp_tc1.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para segmentos com M = 99 amostras.

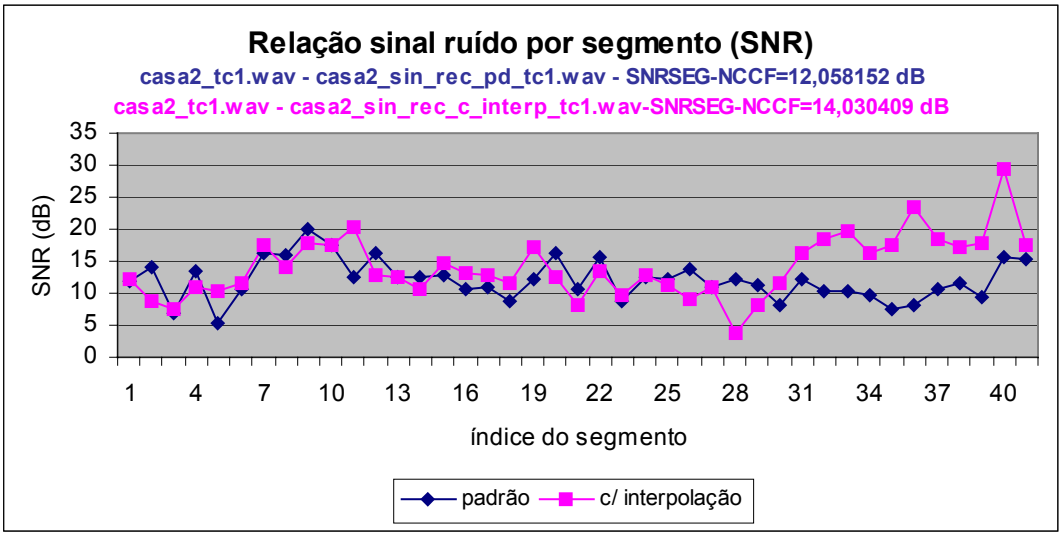

Figura 5.52 – Relação sinal ruído, SNR (dB), entre os segmento no arquivo *casa2_tc1*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd_tc1.wav ou no arquivo casa2_sin_rec_c_interp_tc1.wav. Esses resultados foram obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra SNR (dB) para o sistema de análise – síntese WI (padrão) e na cor magenta mostra SNR (dB) para o sistema de análise – síntese WI (com interpolação).

Na Figura 5.52 Pode-se observar que a SNR para a maioria dos segmentos para o sistema de análise – síntese (com interpolação) apresenta maiores resultados como é mostrado pela curva em magenta. Isto é indicado também pelo valor SNRSEG-NCCF = 14,030409 dB comparada com 12,058152 dB para o sistema WI (padrão). Estes resultados indicam que o sistema de análise – síntese WI (com interpolação) apresenta uma melhor reconstrução do sinal por segmento do que o sistema de análise – síntese WI (padrão) para o sinal do trecho 1.

A Figura 5.53 mostra os deslocamentos d' (em número de amostras) que foram obtidos para cada segmento do arquivo em teste, casa2_sin_rec_pd_tc2.wav ou casa2_sin_rec_c_interp_tc2.wav, em relação ao arquivo original casa2_tc2.wav com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 99 amostras.

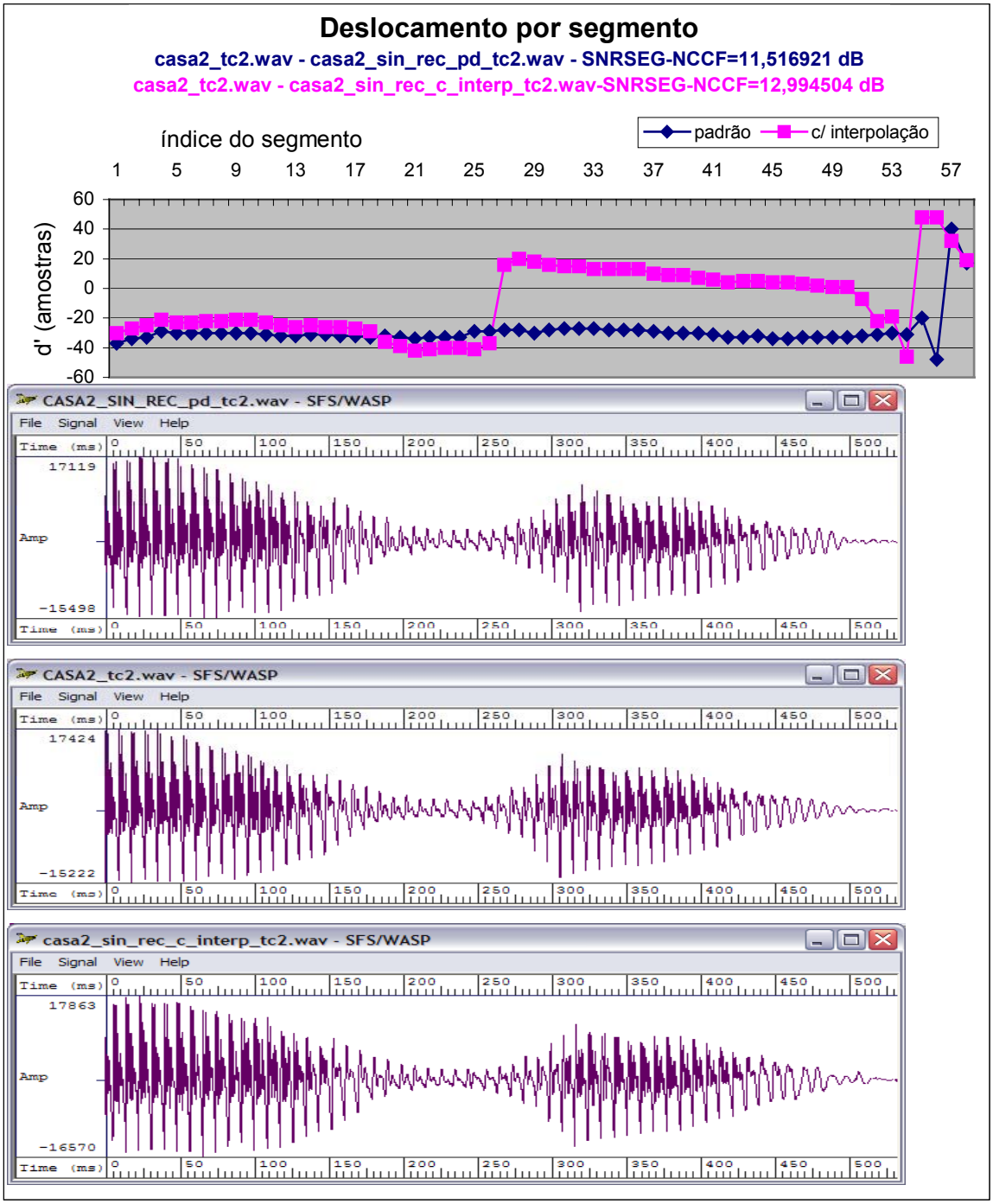

Figura 5.53 – Deslocamentos em número de amostras, d', para os segmentos do sinal reconstruído, no arquivo*casa2*_sin_rec_pd_tc2.wav ou no arquivo casa2_sin_rec_c_interp_tc2.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original, no arquivo *casa2_tc2*.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra d' para o sistema de análise – síntese WI (padrão) e na cor magenta mostra d' para o sistema de análise – síntese WI (com interpolação). Para a visualização da posição relativa dos segmentos são mostrados também os gráficos das formas de onda amplitude x tempo para os sinais da fala nos arquivos envolvidos.

Na Figura 5.53, no primeiro gráfico de cima para baixo, observa-se que o sistema de análise – síntese WI (padrão), na cor azul, apresenta uma seqüência de deslocamentos por segmento mais uniforme. Isto é indicado pelo valor d'med_dif_abs igual a 3 amostras (Tabela 5.9) enquanto que d'med_dif_abs é de 5 amostras (Tabela 5.9) para o o sistema de análise – síntese WI (com interpolação). Para o sistema de análise – síntese WI (com interpolação) observa-se dois trechos onde a seqüência é uniforme mas com algumas descontinuidades. Observa-se então que para os sinais processados no sistema de análise – síntese WI (com interpolação) para o trecho 2 apresentam-se mais defasados do que no sistema de análise – síntese WI (padrão).

A Figura 5.54 mostra as correlações cruzadas normalizadas, R (NCCF), entre os segmentos no arquivo *casa2_tc2*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd_tc2.wav ou no arquivo casa2_sin_rec_c_interp_tc2.wav calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 99 amostras.

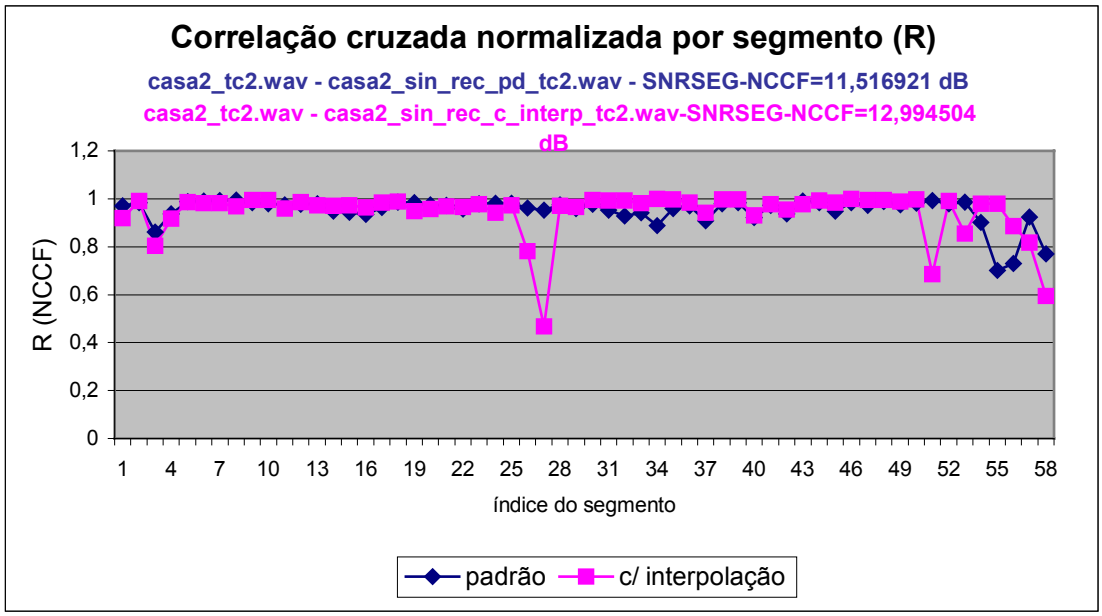

Figura 5.54 – Correlação cruzada normalizada, R (NCCF), entre os segmentos no arquivo *casa2_tc2*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd_tc2.wav ou no arquivo casa2_sin_rec_c_interp_tc2.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra R para o sistema de análise – síntese WI (padrão) e na cor magenta mostra R para o sistema de análise – síntese WI (com interpolação).

Na Figura 5.54 pode-se observar uma seqüência de valores mais uniformes para a correlação R com resultados maiores, por uma pequena margem, para o sistema de análise – síntese WI

(com interpolação) indicados também pela medida SNRSEG-NCCF = 12,994504 dB comparada com 11,516921 dB para o sistema de análise – síntese WI (padrão). Essa Figura mostra também que para o segmento número 28 existe uma descontinuidade que corresponde à mesma posição para a descontinuidade dos deslocamentos indicados na Figura 5.53.

A Figura 5.55 mostra as relações sinal ruído, SNR (dB), entre os segmentos no arquivo *casa2_tc2*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd_tc2.wav ou no arquivo casa2_sin_rec_c_interp_tc2.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 99 amostras.

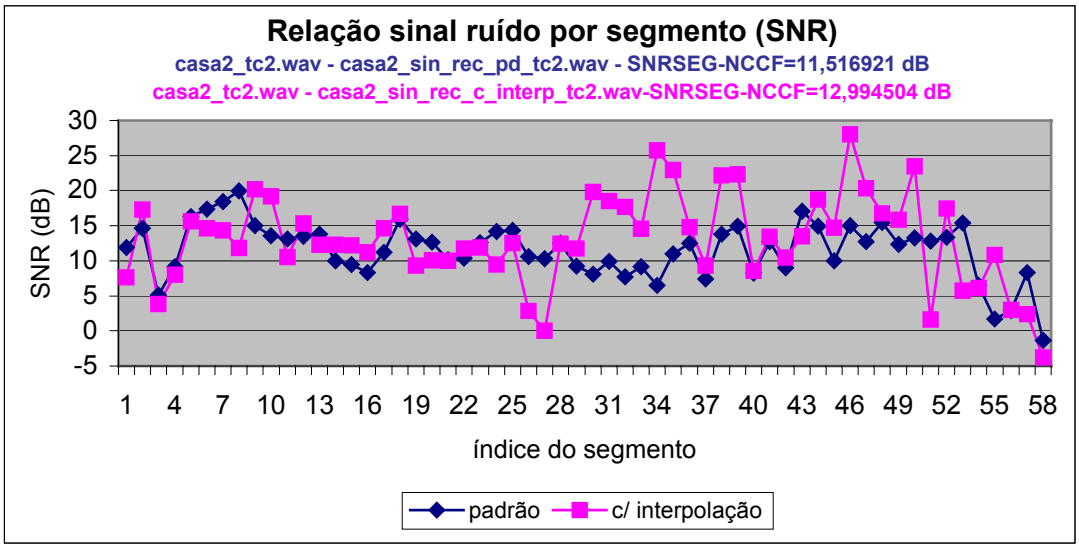

Figura 5.55 – Relação sinal ruído, SNR (dB), entre os segmento no arquivo *casa2_tc2*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd_tc2.wav ou no arquivo casa2_sin_rec_c_interp_tc2.wav. Esses resultados foram obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra SNR (dB) para o sistema de análise – síntese WI (padrão) e na cor magenta mostra SNR (dB) para o sistema de análise – síntese WI (com interpolação).

Na Figura 5.55 pode-se observar que a SNR para a maioria dos segmentos para o sistema de análise – síntese (com interpolação) apresenta maiores resultados como é mostrado pela curva em magenta. Isto também é indicado pelo valor SNRSEG-NCCF = 12,994504 dB e 11,516921 dB para o sistema WI (padrão). Esses resultados indicam que o sistema de análise – síntese WI (com interpolação) apresenta uma melhor reconstrução do sinal por segmento do que o sistema de análise – síntese WI (padrão) para o sinal do trecho 2.

A Figura 5.56 mostra os deslocamentos d' (em número de amostras) que foram obtidos para cada segmento do arquivo em teste, casa2_sin_rec_pd_tc3.wav ou casa2_sin_rec_c_interp_tc3.wav, em relação ao arquivo original casa2_tc3.wav com a aplicação do algoritmo da SNRSEG-NCCF, para segmentos os com M = 99 amostras.

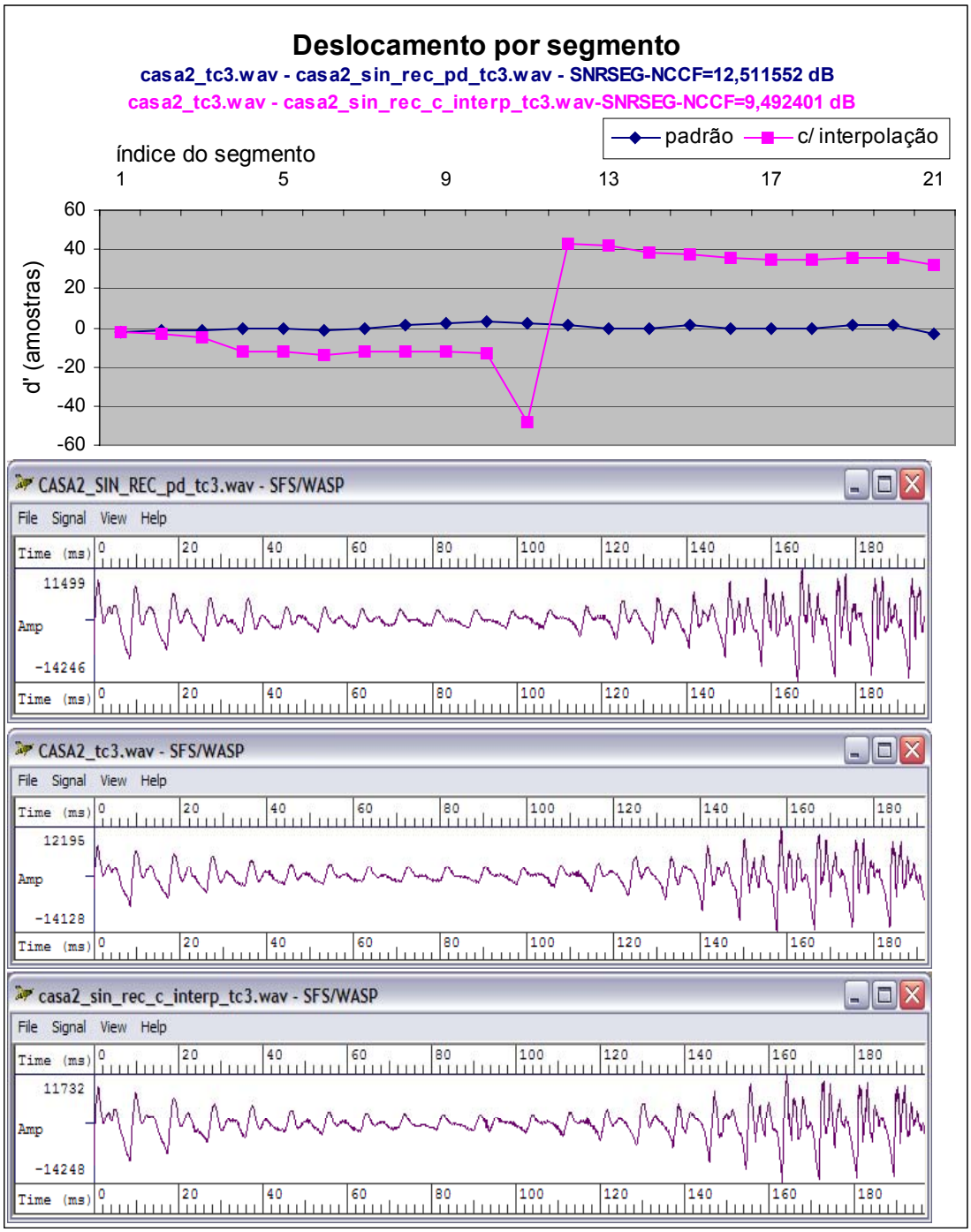

Figura 5.56 – Deslocamentos em número de amostras, d', para os segmentos do sinal reconstruído, no arquivo *casa2*_sin_rec_pd_tc3.wav ou no arquivo casa2_sin_rec_c_interp_tc3.wav, na pesquisa das amostras mais similares às amostras no segmento correspondente no sinal original, no arquivo *casa2_tc3*.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra d' para o sistema de análise – síntese WI (padrão) e na cor magenta mostra d' para o sistema de análise – síntese WI (com interpolação). Para a visualização da posição relativa dos segmentos são mostrados também os gráficos das formas de onda amplitude x tempo para os sinais da fala nos arquivos envolvidos.

Na Figura 5.56, no primeiro gráfico de cima para baixo, observa-se que o sistema de análise – síntese WI (padrão), na cor azul, apresenta uma seqüência de deslocamentos por segmento mais uniforme. Este fato é indicado pelo valor d'med_dif_abs = 0 amostras enquanto que d'med_dif_abs = 7 amostras para o sistema de análise – síntese WI (com interpolação), conforme a Tabela 5.9. Para o sistema de análise – síntese WI (com interpolação) observa-se dois trechos onde a seqüência é uniforme mas com uma descontinuidade entre eles. O primeiro trecho segue de forma aproximada os valores para os deslocamentos do sistema padrão, mas o segundo trecho se afasta da seqüência uniforme em azul. Observa-se então que para os sinais processados no sistema de análise – síntese WI (com interpolação) para o trecho 3 apresentam-se mais defasados que no sistema de análise – síntese WI (padrão).

A Figura 5.57 mostra as correlações cruzadas normalizadas, R (NCCF), entre os segmentos no arquivo *casa2_tc3*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd_tc3.wav ou no arquivo casa2_sin_rec_c_interp_tc3.wav calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 99 amostras.

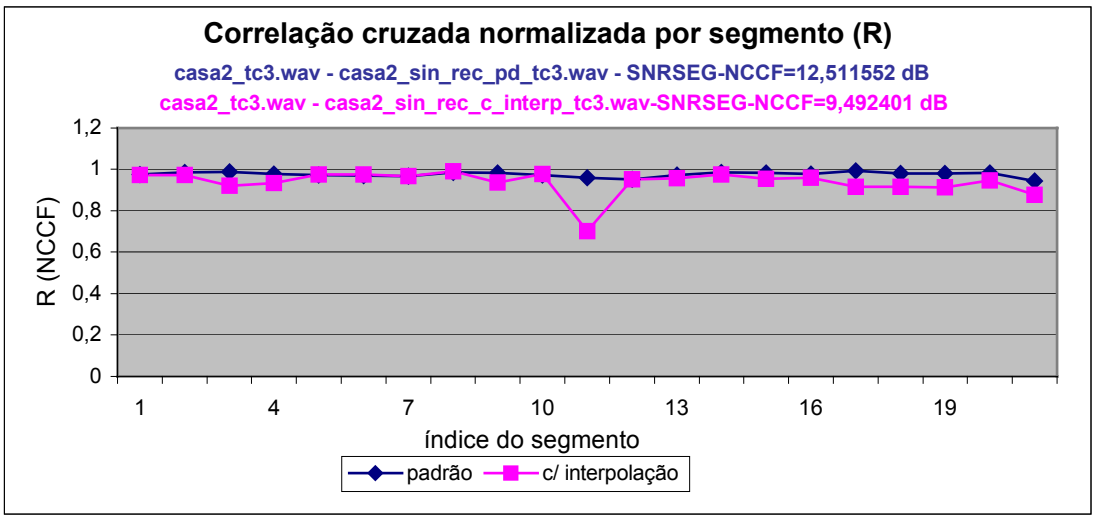

Figura 5.57 – Correlação cruzada normalizada, R (NCCF), entre os segmentos no arquivo *casa2_tc3*.wav, e os segmentos mais similares próximos às posições correspondentes no arquivo *casa2*_sin_rec_pd_tc3.wav ou no arquivo casa2_sin_rec_c_interp_tc3.wav. Resultados obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra R para o sistema de análise – síntese WI (padrão) e na cor magenta mostra R para o sistema de análise – síntese WI (com interpolação).

Na Figura 5.57 observa-se uma seqüência de valores mais uniformes para a correlação R com resultados maiores, por uma pequena margem, para o sistema de análise – síntese WI (padrão) indicados também pela medida SNRSEG-NCCF = 12,511552 dB comparada com 9,492401

dB para o sistema de análise – síntese WI (padrão). Essa Figura mostra também que para o segmento número 11 existe uma descontinuidade que corresponde à mesma posição para a descontinuidade dos deslocamentos indicados na Figura 5.56.

A Figura 5.58 mostra as relações sinal ruído, SNR (dB), entre os segmentos no arquivo *casa2_tc3*.wav e os segmentos correspondentes com a maior similaridade em no arquivo *casa2*_sin_rec_pd_tc3.wav ou no arquivo casa2_sin_rec_c_interp_tc3.wav, calculadas com a aplicação do algoritmo da SNRSEG-NCCF, para os segmentos com M = 99 amostras.

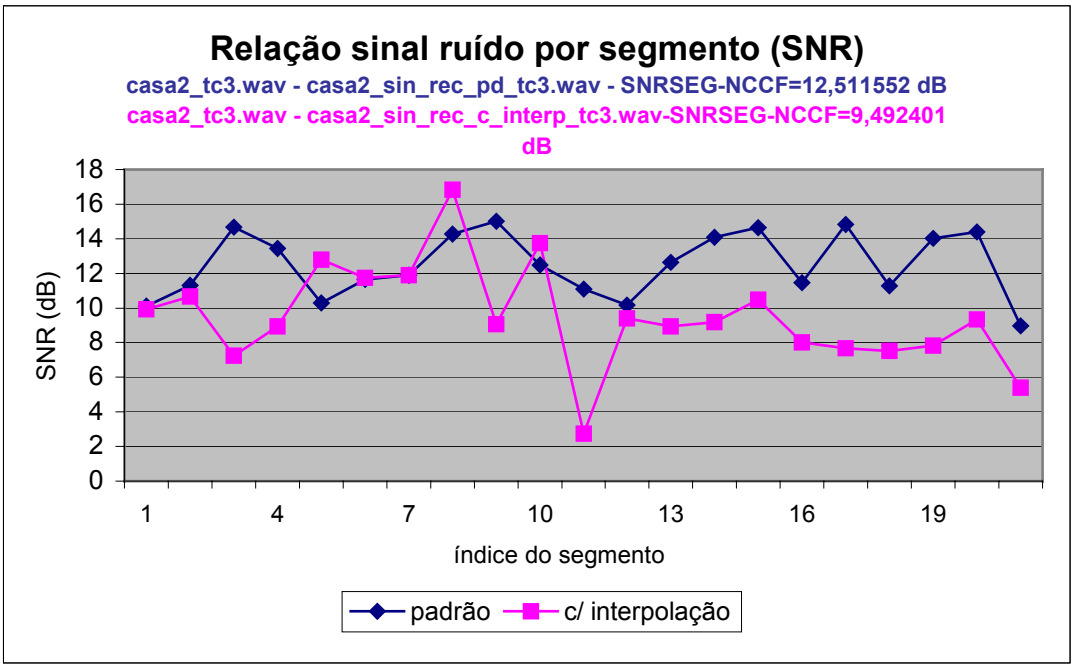

Figura 5.58 – Relação sinal ruído, SNR (dB), entre os segmento no arquivo *casa2_tc3*.wav e os segmentos correspondentes com a maior similaridade no arquivo *casa2*_sin_rec_pd_tc3.wav ou no arquivo casa2_sin_rec_c_interp_tc3.wav. Estes resultados foram obtidos para M = 99 amostras a partir do processamento no algoritmo SNRSEG-NCCF. Na cor azul mostra SNR (dB) para o sistema de analise – síntese WI (padrão) e na cor magenta mostra SNR (dB) para o sistema de análise – síntese WI (com interpolação).

Na Figura 5.58 pode-se observar que a SNR para a maioria dos segmentos para o sistema de análise – síntese (padrão) apresenta maiores resultados como é mostrado pela curva em azul. Isto é indicado também pelo valor SNRSEG-NCCF = 12,511552 dB em comparação com 9,492401 dB para o sistema WI (com interpolação). Estes resultados indicam que o sistema de análise – síntese WI (padrão) apresenta uma melhor reconstrução do sinal por segmento do que o sistema de análise – síntese WI (com interpolação) para o sinal do trecho 3.

Considerações finais (seção 5.2.5.4) **-** Nesta seção foram apresentados os resultados PESQ_MOS e da SNRSEG-NCCF relativos ao processamento dos sinais da fala nos sistemas de análise – síntese WI (padrão) e (com interpolação). O objetivo desta seção foi realizar uma avaliação da versão preliminar do sistema de análise – síntese WI (com interpolação) por meio dos resultados iniciais, comparando os sinais da fala reconstruídos pelos dois sistemas para um mesmo sinal da fala original, utilizando as medidas PESQ_MOS e SNRSEG-NCCF. De forma geral as Tabelas 5.8 e 5.9 e os gráficos subseqüentes apresentam resultados preliminares que permitem concluir que:

- o sistema de análise – síntese WI (padrão) apresenta sinal da fala reconstruído com menor grau de defasagem e com qualidades perceptuais melhores do que o sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação) para sinais completos da fala. Mas em alguns trechos do sinal, o sistema de análise – síntese WI (com interpolação) apresentou melhores qualidades perceptuais para o sinal reconstruído que o sistema de análise – síntese WI (padrão).
- o sistema de análise – síntese WI (com interpolação) apresentou uma reconstrução do sinal da fala por segmento melhor do que o sistema de análise –síntese WI (padrão) que é um indicativo de obtenção de uma forma de onda mais próxima da forma de onda do sinal original;
- Assim existem indicativos de que o sistema de análise – síntese WI (com interpolação) poderá apresentar melhores resultados para a reconstrução global do sinal, a nível de forma de onda e de fase, se houver algum ajuste no modelo. Pode-se citar, por exemplo, a proposta detalhada no Apêndice A desta tese, para ajustar e corrigir o pitch, que poderá resultar em uma melhor reconstrução e em uma menor defasagem entre os sinais originais e reconstruídos.

5.2.5.5 Os resultados: gráficos das formas de onda para o sistema de análise síntese WI (com interpolação)

Nesta seção são apresentados os gráficos das formas de onda (amplitude *versus* tempo) dos sinais originais e dos sinais reconstruídos para o sistema de análise síntese WI (padrão) e para o sistema de análise síntese WI (com interpolação).

5.2.5.5.1 Gráficos das formas de onda (locutor - adulto) - sistema de análise síntese WI (com interpolação)

I - Expressão de fala: "**casa**" (locutor - adulto)

(a) A Figura 5.59 mostra o sinal da fala reconstruído no arquivo ***casa2*_sin_rec_pd.wav,** o sinal da fala original no arquivo ***casa2*.wav**, e o sinal da fala reconstruído no arquivo **casa2_sin_rec_c_interp.wav.** Também são mostradas as localizações para os trechos com os sinais da fala que foram extraídos nas posições correspondentes nos três arquivos.

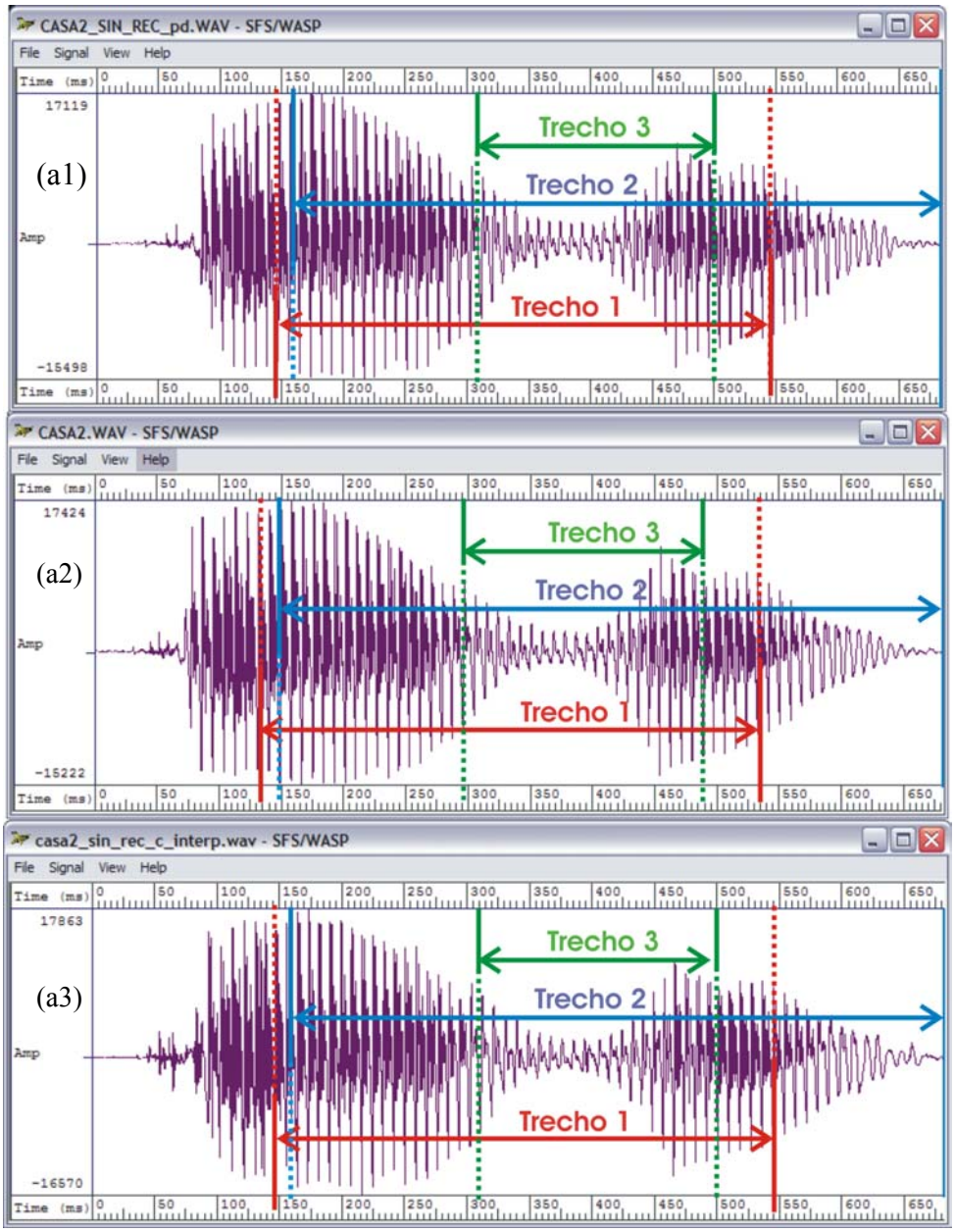

Figura 5.59 – Gráfico das formas de onda (amplitude x tempo) para os sinais (locutor adulto): (a1) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **casa2_sin_rec_pd.wav;** (a2) Sinal da fala original: no arquivo **casa2.wav**; e (a3) Sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **casa2_sin_rec_c_interp.wav**. Em cada gráfico são mostradas as localizações para os trechos onde os sinais da fala foram extraídos nas posições correspondentes nos três arquivos para avaliação com o método PESQ e com o método da SNRSEG-NCCF.

Na Figura 5.59, comparando os gráficos (a1), (a2) e (a3), pode-se observar que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos sinais originais e dos sinais reconstruídos. Para o gráfico (a3): *no início do sinal* para o fonema /c/ observa-se a presença de ruído no sinal; *na posição média do trecho 3* (em torno de 400 ms) existe uma deformação na forma de onda; e *na região final do sinal*, no final do último fonema /a/, observa-se uma pequena diferença entre as amplitudes das formas de ondas.

b) A Figura 5.60 mostra o sinal residual reconstruído da fala no arquivo ***casa2*_res_rec_pd.wav,** o sinal residual original da fala no arquivo ***casa2_res_pd*.wav** e o sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **casa2_res_rec_c_interp.wav**.

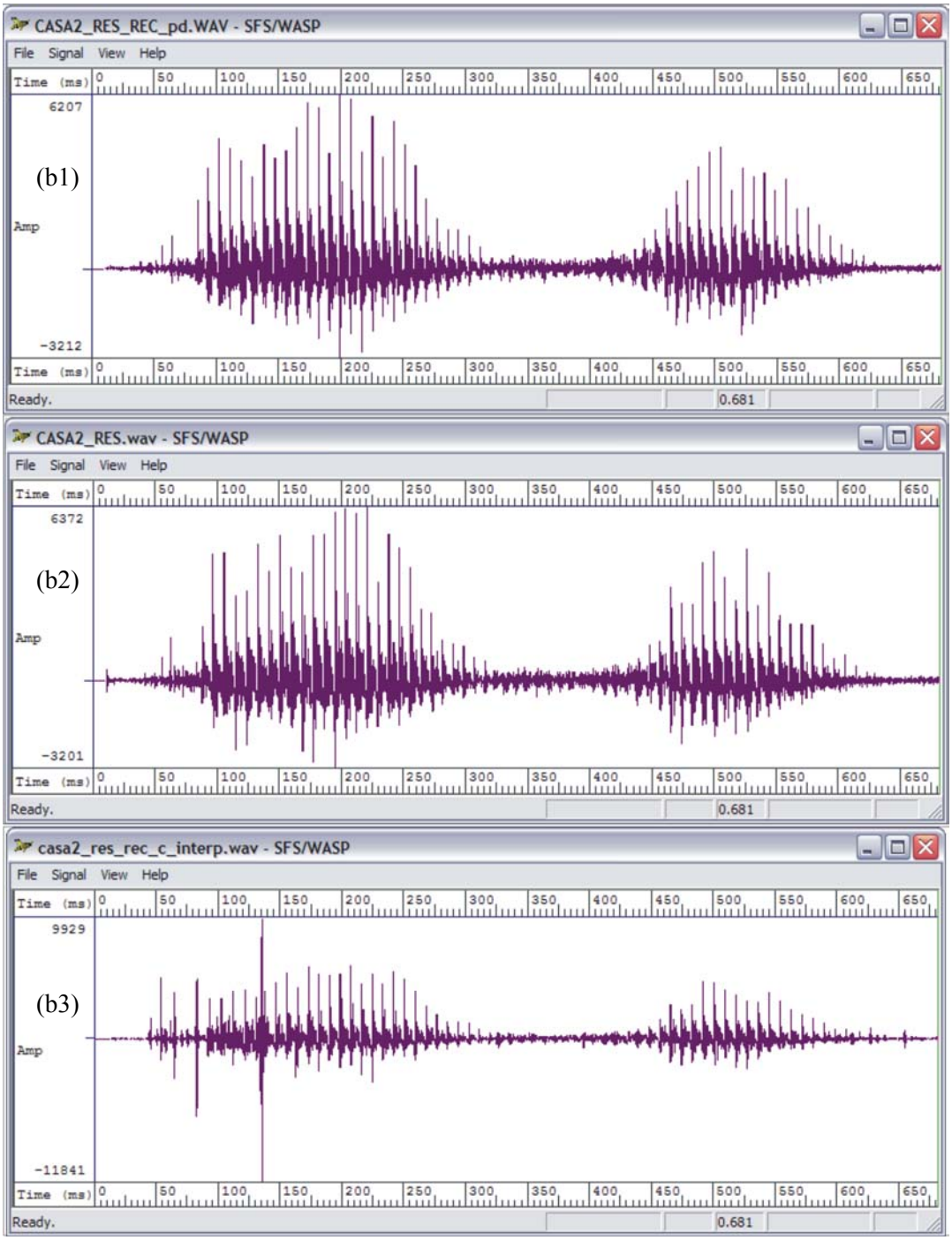

Figura 5.60 – Gráfico das formas de onda (amplitude x tempo) para os sinais residuais (locutor adulto): (b1) Sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **casa2_res_rec_pd.wav;** (b2) Sinal residual original da fala: no arquivo **casa2_res_pd.wav**; e (b3) o sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **casa2_res_rec_c_interp.wav**.

Na Figura 5.60, comparando os gráficos (b1), (b2) e (b3), observa-se que existem semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) do sinal reconstruído pelo sistema (padrão), do sinal original e do sinal

reconstruído pelo sistema (com interpolação). Para o gráfico (b3): *no início do sinal* (até 180 ms) observam-se alguns pontos instáveis no sinal; *a partir de 190 ms* o sinal apresenta-se semelhante ao sinal original (gráfico (b2)).

(c) A Figura 5.61 mostra o sinal da fala reconstruído no arquivo ***casa2*_sin_rec_pd.wav** (na parte superior da figura) e no arquivo **casa2_sin_rec_c_interp.wav** (na parte inferior da figura) e as ampliações de dois trechos dos sinais em posições correspondentes comparados com as ampliações de dois trechos do sinal original no arquivo **casa2.wav** na parte central da figura**.**

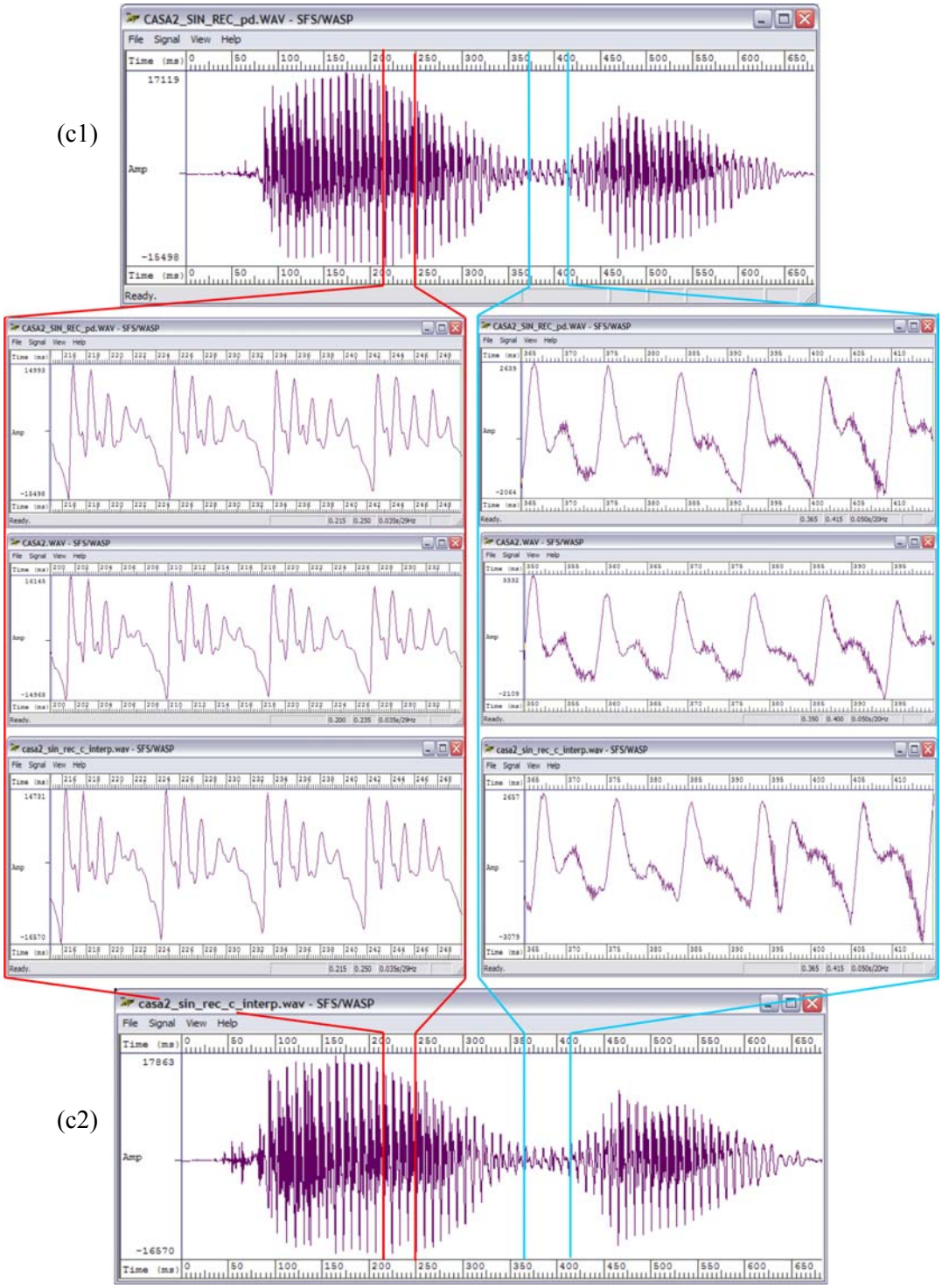

Figura 5.61– Gráfico das formas de onda (amplitude x tempo) para os sinais (locutor adulto): (c1) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **casa2_sin_rec_pd.wav** (correspondente); (c2) Sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **casa2_sin_rec_c_interp.wav** (correspondente); e as ampliações de dois trechos selecionados nas posições correspondentes indicadas nos dois sinais em (c1) e (c2) comparadas com as ampliações de dois trechos correspondentes do sinal original no arquivo **casa2.wav** na parte central da figura**.**

Na Figura 5.61, comparando os gráficos dos trechos ampliados, observa-se que as formas de ondas para o sinal reconstruído pelo sistema (padrão), para o sinal original e para o sinal reconstruído pelo sistema (com interpolação) são semelhantes, exceto para o sinal no último trecho ampliado do gráfico (c2) que apresenta a deformação também observada na Figura 5.59. Esta deformação mostra que provavelmente existe um erro no processo de interpolação com uma contração temporal a ser investigado.

(d) A Figura 5.62 mostra os gráficos das formas de onda (amplitude *versus* tempo) para os sinais localizados pelo trecho 1 de acordo com os gráficos na Figura 5.59: nos arquivos **casa2_sin_rec_pd_tc1.wav**, **casa2_tc1.wav,** e **casa2_sin_rec_c_interp_tc1.wav**.

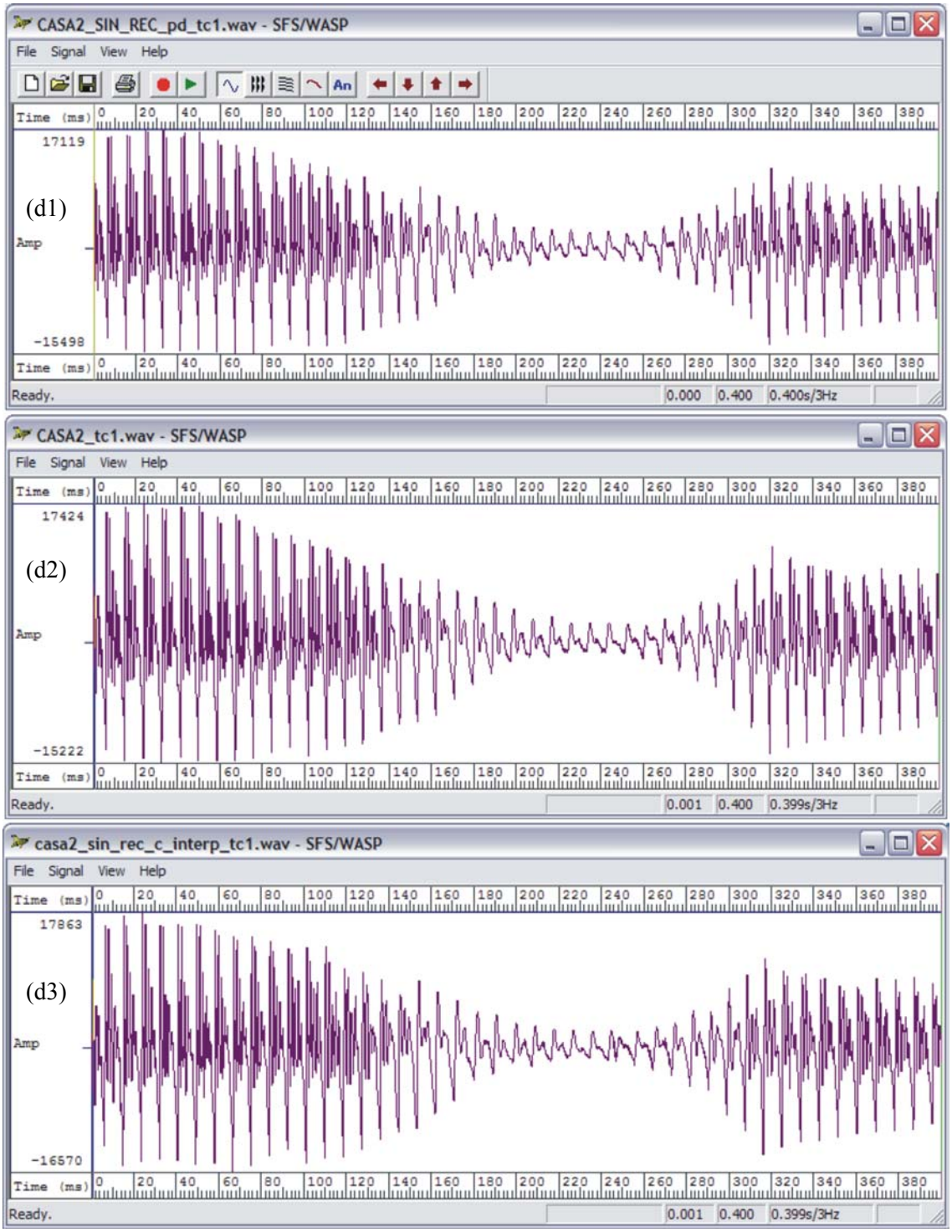

Figura 5.62 - Gráficos das formas de onda (amplitude x tempo) para os sinais localizados pelo trecho 1 de acordo com os gráficos da Figura 5.59. (d1) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **casa2_sin_rec_pd_tc1.wav**; (d2) Sinal da fala original: no arquivo **casa2_tc1.wav;** e (d3) Sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **casa2_sin_rec_c_interp_tc1.wav**. Os trechos com os sinais da fala foram extraídos nas posições correspondentes nos arquivos **casa2_sin_rec_pd.wav, casa2.wav e casa2_sin_rec_c_interp.wav,** respectivamente.

Na Figura 5.62, comparando os gráficos (d1), (d2) e (d3), observa-se que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos trechos dos sinais originais e dos sinais reconstruídos. Também pode-se observar que as formas de onda para o trecho do sinal reconstruído pelo sistema (padrão), para o trecho do sinal original e para o trecho do sinal reconstruído pelo sistema (com interpolação) são semelhantes, exceto para o trecho do sinal do gráfico (d3) que apresenta a deformação também observada nas Figuras 5.59 e 5.61.

(e) A Figura 5.63 mostra os gráficos das formas de onda (amplitude *versus* tempo) para os sinais localizados pelo trecho 2 de acordo com os gráficos na Figura 5.59: nos arquivos **casa2_sin_rec_pd_tc2.wav**, **casa2_tc2.wav,** e **casa2_sin_rec_c_interp_tc2.wav**.

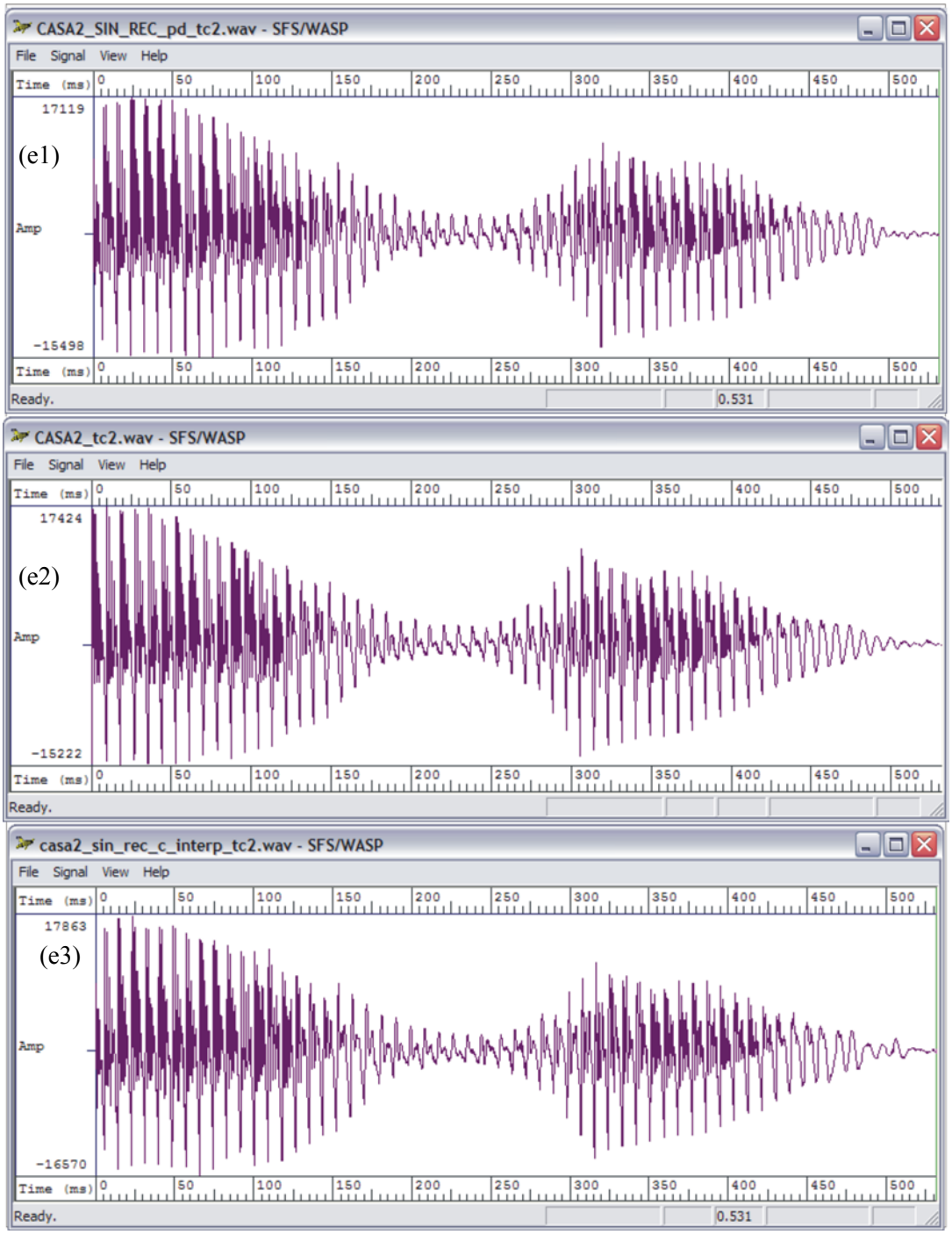

Figura 5.63 - Gráficos das formas de onda (amplitude x tempo) para os sinais localizados pelo trecho 2 de acordo com os gráficos na Figura 5.59. (e1) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **casa2_sin_rec_pd_tc2.wav**; (e2) Sinal da fala original: no arquivo **casa2_tc2.wav;** e (e3) Sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **casa2_sin_rec_c_interp_tc2.wav**. Os trechos com os sinais da fala foram extraídos nas posições correspondentes nos arquivos **casa2_sin_rec_pd.wav, casa2.wav e casa2_sin_rec_c_interp.wav,** respectivamente.

Na Figura 5.63, comparando os gráficos (e1), (e2) e (e3), observa-se que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos trechos dos sinais originais e dos sinais reconstruídos. Também se pode-se observar que as formas de onda são semelhantes para o trecho do sinal reconstruído pelo sistema (padrão), para o trecho do sinal original e para o trecho do sinal reconstruído pelo sistema (com interpolação), exceto para o trecho do sinal do gráfico (e3) que apresenta a deformação também observada nas Figuras 5.59, 5.61 e 5.62. Na *região final do sinal*, no final do último fonema /a/, observa-se uma pequena diferença entre as amplitudes das formas de onda que também já foi verificada na Figura 5.59.

(f) A Figura 5.64 mostra os gráficos das formas de onda (amplitude *versus* tempo) para os sinais localizados pelo trecho 3 de acordo com os gráficos na Figura 5.59: nos arquivos **casa2_sin_rec_pd_tc3.wav**, **casa2_tc3.wav,** e **casa2_sin_rec_c_interp_tc3.wav**.

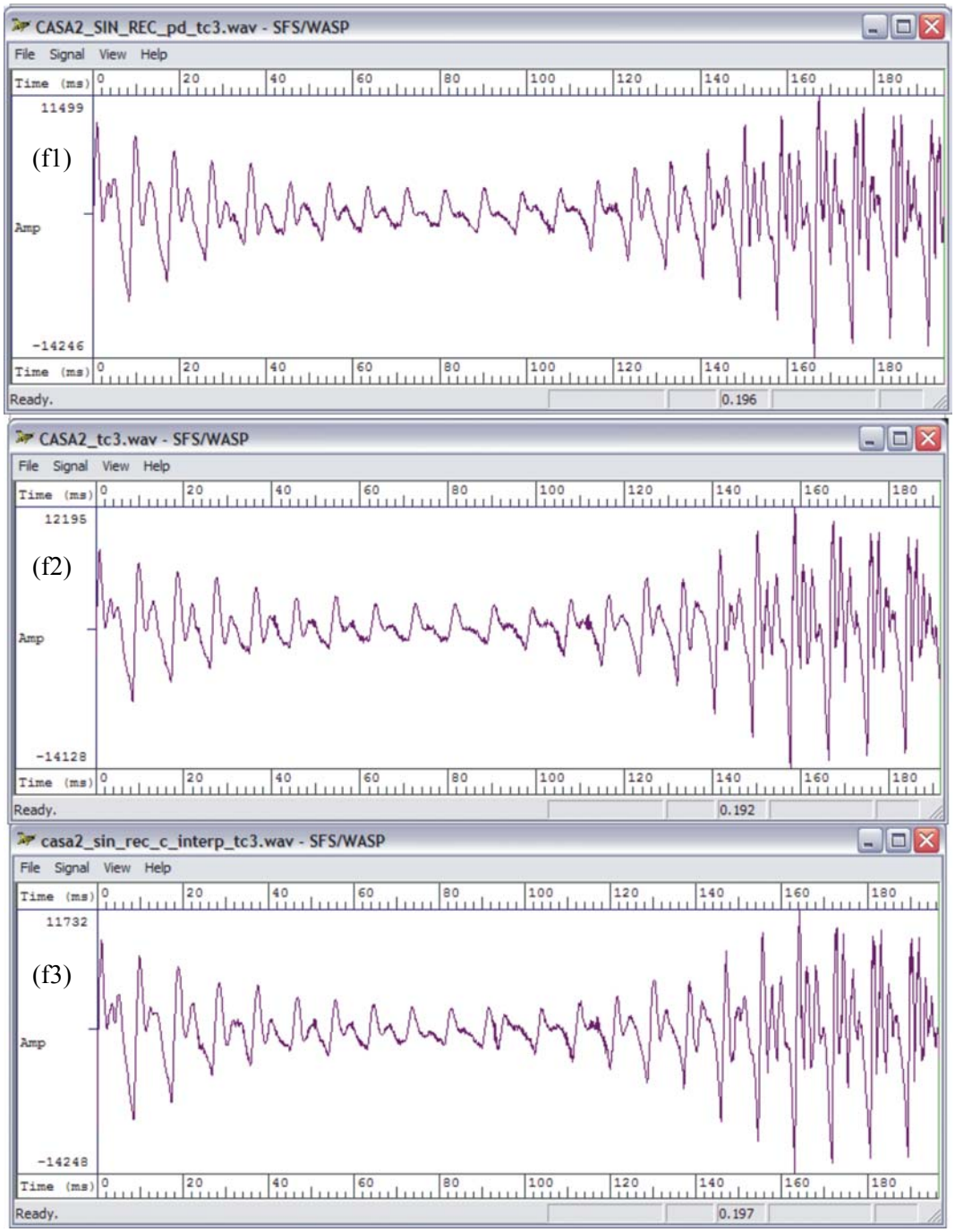

Figura 5.64 - Gráficos das formas de onda (amplitude x tempo) para os sinais localizados pelo trecho 2 de acordo com os gráficos na Figura 5.59. (f1) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **casa2_sin_rec_pd_tc3.wav**; (f2) Sinal da fala original: no arquivo **casa2_tc3.wav;** e (f3) Sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **casa2_sin_rec_c_interp_tc3.wav**. Os trechos com os sinais da fala foram extraídos nas posições correspondentes nos arquivos **casa2_sin_rec_pd.wav, casa2.wav e casa2_sin_rec_c_interp.wav,** respectivamente.

Na Figura 5.64, comparando os gráficos (f1), (f2) e (f3), pode-se observar que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos trechos dos sinais originais e dos sinais reconstruídos. Também se observa que as formas de onda para o trecho do sinal reconstruído pelo sistema (padrão), para o trecho do sinal original e para o trecho do sinal reconstruído pelo sistema (com interpolação) são semelhantes, exceto para o trecho do sinal do gráfico (f3) que apresenta a deformação também observada nas Figuras 5.59, 5.61, 5.62 e 5.63.

II - Expressão de fala: "**é bonita**" (locutor - adulto)

(a) A Figura 5.65 mostra o sinal da fala original ***ebonita2*.wav**, os sinais da fala reconstruídos ***ebonita2*_sin_rec_pd.wav** e **ebonita2_sin_rec_c_interp.wav.**

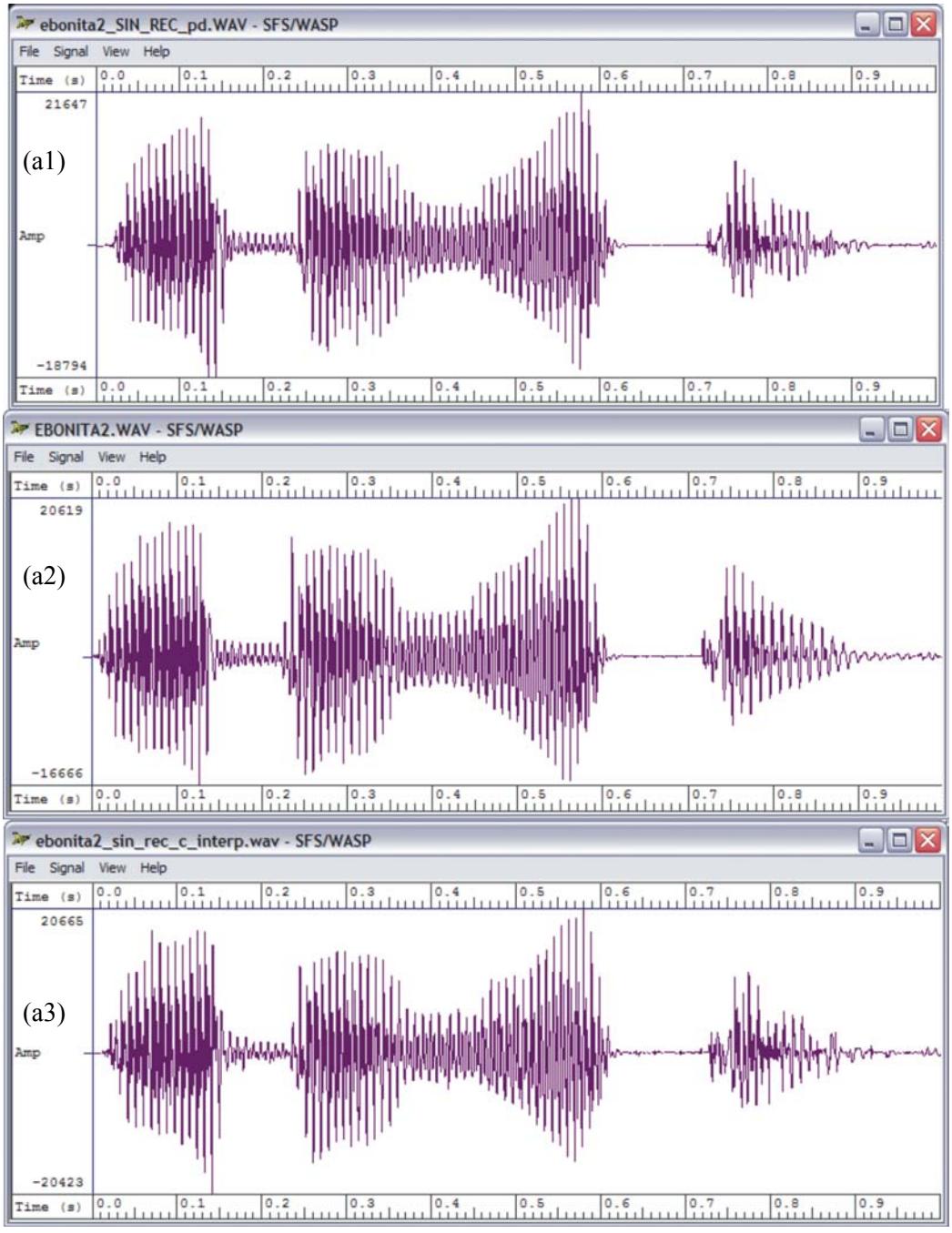

Figura 5.65 - Gráfico das formas de onda (amplitude x tempo) para os sinais (locutor adulto): (a1) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **ebonita2_sin_rec_pd.wav** (correspondente)**;** (a2) Sinal da fala original: no arquivo **ebonita2.wav**; e (a3) Sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **ebonita2_sin_rec_c_interp.wav** (correspondente).

Na Figura 5.65, comparando os gráficos (a1), (a2) e (a3), pode-se observar que existem grandes semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes máximas e mínimas) dos trechos dos sinais originais e dos sinais reconstruídos. Para o último fonema /a/, na região final do sinal, nota-se a diferença entre as amplitudes das

formas de onda tanto no gráfico (a1) para o sinal reconstruído pelo sistema de análise – síntese WI (padrão) quanto no gráfico (a3) para o sinal reconstruído pelo sistema de análise – síntese WI (com interpolação) onde a deformação foi mais intensa.

b) A Figura 5.66 mostra o sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo ***ebonita2*_res_rec_pd.wav**, o sinal residual original da fala no arquivo ***ebonita2_res_pd*.wav** e o sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **ebonita2_res_rec_c_interp.wav**.

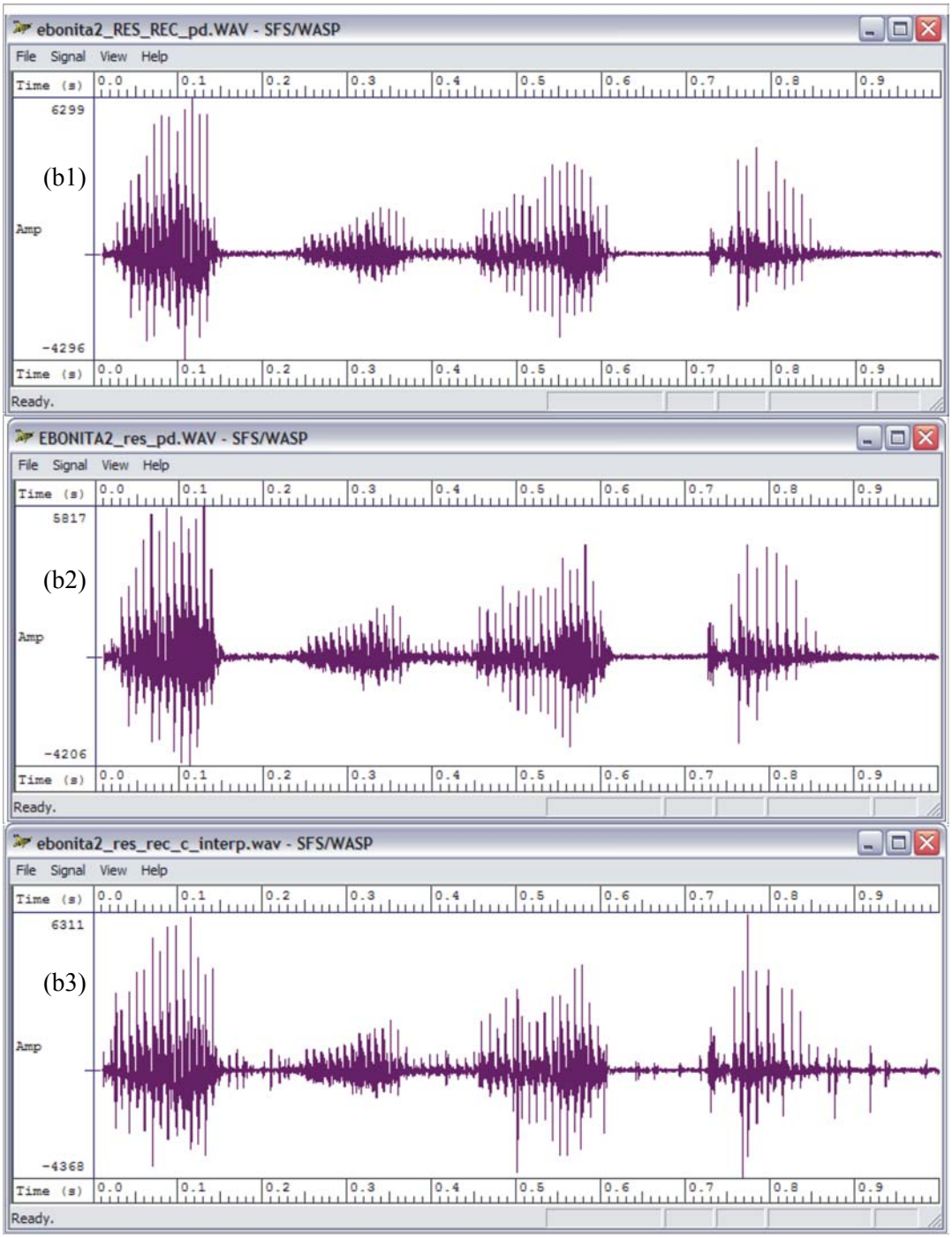

Figura 5.66 - Gráfico das formas de onda (amplitude x tempo) para os sinais residuais (locutor adulto): (b1) Sinal residual da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **ebonita2_res_rec_pd.wav** (correspondente)**;** (b2) Sinal residual original da fala: no arquivo **ebonita2_res_pd.wav**; e (b3) o sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **ebonita2_res_rec_c_interp.wav** (correspondente).

Na Figura 5.66, comparando os gráficos (b1), (b2) e (b3), pode-se observar que existem semelhanças entre o formato geral da envoltória temporal (contorno com as amplitudes

máximas e mínimas) do sinal reconstruído pelo sistema (padrão), do sinal original e do sinal reconstruído pelo sistema (com interpolação). Para o gráfico (b3) observa-se alguns pontos com instabilidade no sinal. Esta deformação (a ser investigada) provavelmente apareceu devido ao processo de interpolação do sistema de análise – síntese WI (com interpolação).

c) A Figura 5.67 mostra o sinal da fala reconstruído no arquivo **ebonita2_sin_rec_pd.wav** (na parte superior) e no arquivo **ebonita2_sin_rec_c_interp.wav** (na parte inferior) e as ampliações de dois trechos dos sinais nas posições correspondentes comparados com as ampliações de dois trechos do sinal original no arquivo **ebonita2.wav** na parte central da figura**.**

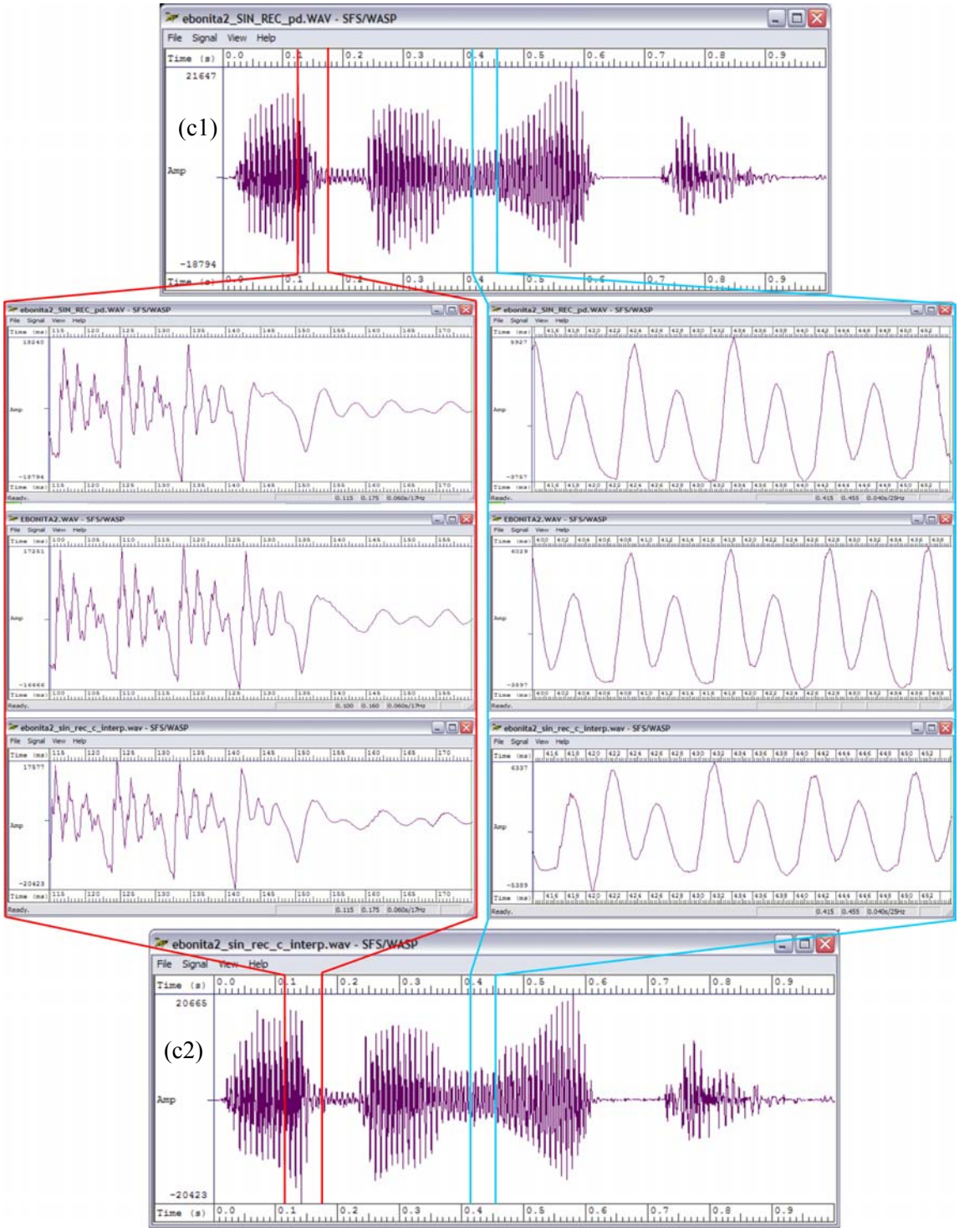

Figura 5.67– Gráfico das formas de onda (amplitude x tempo) para os sinais (locutor adulto): (c1) Sinal da fala reconstruído pelo sistema de análise – síntese WI (padrão): no arquivo **ebonita2_sin_rec_pd.wav** (correspondente); (c2) Sinal da fala reconstruído pelo sistema de análise – síntese WI (com interpolação): no arquivo **ebonita2_sin_rec_c_interp.wav** (correspondente); e as ampliações de dois trechos selecionados nas posições correspondentes indicadas nos dois sinais em (c1) e (c2) comparadas com as ampliações de dois trechos correspondentes do sinal original no arquivo **ebonita2.wav** na parte central da figura**.**

Na Figura 5.67, comparando os gráficos dos trechos ampliados, observa-se que as formas de onda para o sinal reconstruído pelo sistema (padrão), para o sinal original e para o sinal reconstruído pelo sistema (com interpolação) são semelhantes.

Considerações finais (seção 5.2.5.5) - Nesta seção foram apresentados os gráficos das formas de onda (amplitude x tempo) dos sinais originais e dos sinais reconstruídos para o sistema de análise síntese WI (padrão) e para o sistema de análise síntese WI (com interpolação). Foram considerados os gráficos dos sinais utilizados na avaliação da versão preliminar do sistema de análise – síntese WI (com interpolação) comparados com os gráficos dos sinais processados no sistema de análise – síntese WI (padrão). Para esta avaliação foram considerados os processamentos dos sinais da fala nos arquivos casa2.wav e ebonita2.wav relativos ao locutor adulto.

De forma geral os gráficos dos sinais reconstruídos para os sinais da fala e para os sinais residuais da fala nos dois sistemas mostraram-se semelhantes em relação à envoltória temporal (contorno com as amplitudes máximas e mínimas) do sinal e também em relação às formas de onda em relação aos gráficos dos sinais originais. Ocorreram algumas exceções :

- em relação à envoltória temporal os dois sistemas apresentaram deformações na amplitude do sinal reconstruído para o fonema /a/ na expressão "é bonita";
- Para o fonema /z/ na expressão "casa" o sistema de análise – síntese WI (com interpolação) apresentou uma deformação na forma de onda que parece estar relacionada com o processo de interpolação que deverá ser investigada na continuação desta pesquisa.

Em geral o sistema de análise – síntese WI (com interpolação) com ajustes no processo de interpolação mostra que tem potencial para proporcionar uma melhor reconstrução do sinal e uma melhor qualidade da fala que o sistema de análise – síntese WI (padrão).

5.3 Considerações finais deste capítulo

Este capítulo apresentou os resultados das simulações que mostraram o desempenho do sistema de análise – síntese WI (padrão) e o desempenho da versão preliminar do sistema de análise – síntese WI (com interpolação).

Para fazer a avaliação foram utilizados os métodos PESQ efetuando a medida PESQ_MOS e o método da SNRSEG-NCCF, desenvolvido neste trabalho, calculando as medidas SNRSEG-NCCF, d'med e d'med_dif_abs. A medida PESQ_MOS foi utilizada para fazer a avaliação da qualidade perceptual dos sinais da fala reconstruídos pelos sistemas de análise – síntese WI. A medida SNRSEG-NCCF foi utilizada para avaliar a reconstrução da forma de onda dos sinais reconstruídos por segmento e a medida d'med_dif_abs foi utilizada para avaliar a defasagem dos sinais reconstruídos por segmento. A medida d'med (ou

defasagem média) foi utilizada como uma medida auxiliar durante a execução do algoritmo da SNRSEG-NCCF e também para os trechos dos sinais sonoros ou trechos dos sinais residuais correspondentes, o deslocamento médio d'méd foi útil para a comparação com a *defasagem média* obtida durante a inspeção visual das formas de ondas, nos gráficos da amplitude *versus* tempo, na verificação da defasagem entre os sinais.

Para a avaliação dos sistemas foram utilizados os sinais da fala gravados nos arquivos ".wav" usando as expressões da fala "casa" e "é bonita" que foram pronunciadas por uma locutora adulta, uma locutora infantil e um locutor adulto.

Para avaliar os sistemas foram considerados os seguintes resultados:

I – Avaliação do sistema de análise – síntese WI (padrão):

(a) avaliação dos sinais reconstruídos da fala e dos sinais residuais da fala dos três locutores pelas medidas PESQ_MOS e SNRSEG-NCCF;

(b) avaliação da defasagem pelas medidas d'med_dif_abs e pela inspeção visual das formas de onda: entre os sinais reconstruídos da fala e os sinais originais da fala; e entre os sinais residuais reconstruídos e os sinais residuais originais da fala;

(c) avaliação da reconstrução dos sinais da fala e dos sinais residuais da fala pelos gráficos das formas de onda (amplitude *versus* tempo).

II – Avaliação da versão preliminar do sistema de análise – síntese WI (com interpolação):

(a) avaliação dos sinais reconstruídos da fala e dos sinais residuais da fala do locutor adulto pelas medidas PESQ_MOS e SNRSEG-NCCF para a expressão da fala "casa" e "é bonita" em comparação as mesmas medidas para o sistema de análise – síntese WI (padrão);

(b) avaliação dos sinais reconstruídos da fala e dos sinais residuais da fala do locutor adulto pelas medidas PESQ_MOS e SNRSEG-NCCF para a expressão da fala "casa" e alguns trechos no sinal "casa" em comparação com as mesmas medidas para o sistema de análise – síntese WI (padrão): pelas medidas PESQ_MOS e SNRSEG-NCCF e dos gráficos para o deslocamento, NCCF e SNR por segmento.

Os resultados e as análises ao longo do capítulo, mostraram que em algumas simulações a versão preliminar do sistema de análise – síntese WI (com interpolação) apresentou melhor desempenho que o sistema de análise – síntese WI (padrão) na reconstrução das formas de onda por segmento, mas houve um aumento na defasagem dos sinais por segmento sendo portanto necessário efetuar ajustes no processo de interpolação que poderão melhorar o desempenho do sistema. Como exemplo, o sistema pode ser melhorado com o processo de ajuste do pitch proposto no Apêndice A desta tese.

Também os resultados e as análises mostraram de forma geral que o sistema de análise – síntese WI (padrão) apresentou uma boa reconstrução da forma de onda, da envoltória temporal indicando bom controle sobre a potência do sinal e bons resultados para a fala

reconstruída com qualidade perceptual entre as faixas "satisfatória" e "boa" (de acordo com a Tabela 1.1 e Figura 1.1 no capítulo 1) com alguns sinais próximos da "Toll Quality".

Os resultados também confirmam, conforme Choy [14], que o sistema de análise – síntese WI descrito no capítulo 4 deste trabalho, de forma geral, não mantém o sincronismo temporal entre o sinal original da fala e o sinal reconstruído da fala. Isto ocorre principalmente devido a imprecisão na estimação do caminho do pitch e do "time scaling process" durante a interpolação das formas de onda no processo D230. Entretanto a sincronização no tempo poderá ser atingida (ou perseguida) se forem obedecidas as seguintes condições: - executar a extração da CW para cada posição de amostragem ou para cada período de amostragem; - obter um exato caminho da fase $\phi(n)$; - obter o valor exato para a fase inicial $\phi(0)$; - durante o processo de extração das CW's, não permitir o deslocamento da janela de extração fazendo o fator $\varepsilon = 0$; - e preservar a componente DC em cada CW, primeiro coeficiente da DTFS. Na prática é muito difícil, ou impossível de se conseguir o valor exato para o pitch a cada ponto de amostragem do sinal da fala. Desta forma este sistema tem uma implementação mais próxima da implementação do sistema descrito na referência [63] que afirma: "Mesmo em altas taxas de amostragem das CW's o sistema não mantém o sincronismo entre o sinal original e o sinal reconstruído". Enquanto esta implementação está mais distante do sistema descrito por Miguel Arjona [68] que tem uma maior restrição com os problemas de sincronia e apresenta maior sincronismo entre os sinais originais e os sinais recontruídos da fala.

Capítulo 6

CONCLUSÃO

6.1 Introdução

Neste capítulo são apresentadas as conclusões e as contribuições relativas a tese e uma proposta para trabalhos futuros.

6.2 Conclusões

O Sistema de Análise - Síntese WI (padrão)

O objetivo principal desta tese foi desenvolver parte de um sistema de codificação da fala por interpolação de ondas utilizando a técnica WI, ou seja, um codificador/decodificador WI operando na camada análise-síntese. Este tipo de sistema não tem algoritmo disponível na literatura científica, pois a técnica WI é nova e promissora, o que resulta em grandes interesses comerciais entre as companhias de telecomunicações e os grupos de pesquisadores envolvidos. As publicações relativas ao sistema WI geralmente são apresentadas em um certo nível que não permite uma implementação direta do sistema WI. O grupo de pesquisa ou pesquisador interessado nesta área normalmente tem que desenvolver o seu próprio sistema WI. O trabalho desta tese possibilita o início do desenvolvimento do sistema codificador/decodificador WI para este grupo de pesquisa, ao desenvolver uma parte do sistema WI.

Neste trabalho foi desenvolvida a camada externa do sistema WI, ou camada de análise - síntese WI, ficando para futuras pesquisas o desenvolvimento da camada interna do sistema, ou camada de compressão dos parâmetros para compor o sistema codificador/decodificador WI. A camada externa desenvolvida foi então denominada de *sistema de análise – síntese WI (padrão)*.

Na tese foi apresentada uma visão geral da técnica de codificação da fala por interpolação de ondas, a técnica WI e a estrutura básica para um codificador/decodificador utilizando a técnica WI, sem o uso da compressão dos parâmetros, ou seja, o sistema denominado de *sistema de análise – síntese WI (padrão)*. O objetivo foi descrever com detalhes os processos relativos à camada de análise-síntese envolvidos no sistema codificador/decodificador para facilitar o entendimento e permitir a visualização de um algoritmo para a implementação de um simulador para o sistema codificador/decodificador

WI operando na camada de análise-síntese. São apresentados detalhes dos processos e dos procedimentos envolvidos e alguns algoritmos que não estão disponíveis na literatura.

A partir deste trabalho e auxiliado pelas informações em [14], foi possível implementar, em linguagem C, um simulador para o codificador WI na camada de análise-síntese, que tornou-se a parte principal deste trabalho. A visualização e a implementação do algoritmo para o analisador/sintetizador WI foram necessárias, pois o algoritmo não está disponível na literatura científica. Com a implementação e os resultados obtidos ficou caracterizado que a partir das informações deste trabalho, das informações em [14] e das referências bibliográficas é possível implementar um codificador/decodificador WI básico e convencional, ou seja, um *sistema de análise-síntese WI (padrão),* onde *básico* significa operando na camada de análise – síntese, e *convencional* significa sinal reconstruído sem reconstrução perfeita.

O Sistema de Análise – Síntese WI (com interpolação)

Também neste trabalho foi propiciado o início do aperfeiçoamento do sistema de análise – síntese WI (padrão) com a implementação e a avaliação de uma versão preliminar do sistema de análise – síntese WI com melhoria no processo de extração das CW's no lado da análise, pela interpolação das CW's em posições regulares, com o objetivo de conseguir um sinal reconstruído mais similar ao sinal original da fala. Este sistema foi denominado de *sistema de análise – síntese WI (com interpolação)*. O sistema foi apresentado de forma resumida e a avaliado na Seção 5.2.5. E, no apêndice A são mostrados os detalhes deste sistema.

A avaliação dos sistemas de análise – síntese WI

No capítulo 5 foram apresentados os resultados obtidos a partir das simulações com alguns sinais da fala, que mostram o desempenho dos sistemas implementados neste trabalho, o sistema de análise-síntese WI (padrão) e a versão preliminar do sistema de análise – síntese WI (com interpolação). Na avaliação foram utilizados os métodos PESQ, como avaliação objetiva da qualidade perceptual da fala, e o método SNRSEG - NCCF desenvolvido neste trabalho, como avaliação objetiva da reconstrução das formas de ondas dos sinais por segmento, e da defasagem por segmento entre os sinais originais e reconstruídos. Ainda completando, com uma avaliação visual (subjetiva) foram mostrados e comparados os gráficos (amplitude x tempo) das formas de onda dos sinais da fala utilizados nas simulações, com os sinais originais da fala e os sinais reconstruídos da fala, e também, os sinais residuais originais (após a análise) e os sinais residuais reconstruídos (antes da síntese).

Avaliação da versão preliminar do sistema de análise – síntese WI (com interpolação) - Este sistema é uma versão aperfeiçoada do sistema de análise-síntese WI (padrão) com a aplicação parcial da proposta apresentada no apêndice A desta tese. Conforme os resultados mostrados no capítulo 5, a versão preliminar do sistema de análise – síntese WI (com interpolação) ainda não apresentou o desempenho esperado pela proposta. As medidas SNRSEG-NCCF indicam uma melhor *reconstrução das formas de onda por segmento* neste sistema em relação ao sistema de análise – síntese WI (padrão). Também para alguns trechos, ou parte do sinal da fala verificado, o sistema apresentou a medida PESQ_MOS superior, mas com uma maior *defasagem por segmento* que o sistema de análise – síntese WI (padrão). Assim, pode-se dizer que o processo de interpolação das CW's em posições regulares, de acordo com o modelo proposto e implementado, apresentou melhor reconstrução do sinal da fala por segmento, mas também ocorreu um aumento no grau de defasagem entre os sinais originais e reconstruídos. Com esta versão preliminar o sistema de análise – síntese WI (com interpolação) não apresentou de forma geral uma melhora na qualidade da fala reconstruída. Porém, os resultados indicam que o modelo apresentado tem potencial para apresentar o sinal reconstruído da fala com melhor qualidade perceptual se forem efetuados ajustes no processo de interpolação para diminuir a defasagem dos sinais. Um exemplo seria executar o ajuste do pitch proposto no apêndice A.

Avaliação do sistema de análise – síntese WI (padrão) - Conforme os resultados, o sistema apresentou boa reconstrução do sinal indicado pela similaridade entre os envelopes temporais e entre as formas de ondas (que mostram um bom controle sobre a potência do sinal), e pelas medidas SNRSEG-NCCF entre os sinais avaliados, os sinais originais e os sinais reconstruídos da fala. O parâmetro d'med_dif_abs mostrou que este sistema também apresentou um grau de defasagem menor do que a defasagem apresentada pelo sistema de análise – síntese WI (com interpolação). As medidas PESQ_MOS indicam que a qualidade perceptual da fala reconstruída, conforme a classificação MOS (Tabela 1.1 e Figura 1.1 no capítulo 1), ficou entre "satisfatória" e "boa" situada na faixa "communication quality". A análise da defasagem entre os sinais originais e reconstruídos pelo sistema de análise – síntese WI (padrão) mostram que os sinais em geral não estão sincronizados no tempo [14]. Para trechos ruidosos, região de transição entre /fonemas consoantes/ e /fonemas vogais/ têm defasagem variável. Trechos sonoros mais potentes apresentam regiões com defasagem quase constante e regiões com defasagem constante. Deste modo, os resultados confirmam que o sistema de análise – síntese WI descrito no capítulo 4 deste trabalho, de forma geral, não mantém o sincronismo temporal entre o sinal original da fala e reconstruído da fala conforme Choy [14], estando este sistema com uma implementação mais próxima da implementação do sistema descrito por Deep Sen [63] do que do sistema descrito por Miguel Arjona [68], que

tem uma maior restrição com os problemas de sincronia e apresenta maior sincronismo entre os sinais originais e os reconstruídos da fala.

O sistema de análise – síntese WI fica disponível no grupo de pesquisa para continuidade as pesquisas com o sistema codificador/decodificador WI. De forma geral o sistema está pronto para possibilitar a continuidade da pesquisa com a implementação da camada interna, ou camada de compressão dos parâmetros, completando o codificador/decodificador WI ou sistema WI.

6.3 Contribuições

De forma geral este trabalho contribui com a apresentação e a descrição da técnica WI, aplicada ao longo da descrição dos processos que compõem um sistema de codificação da fala (codificador/decodificador básico) com detalhes dos procedimentos envolvidos e alguns algoritmos de forma que ajuda a esclarecer e a complementar parte do trabalho desenvolvido por Choy [14], para permitir uma melhor visualização de um algoritmo para implementar o codificador/decodificador WI – básico (sem compressão dos parâmetros).

Na apresentação e na descrição procurou-se fornecer um enfoque didático, uma visão global do sistema e ao mesmo tempo uma visão específica, detalhada e resumida dos processos enfatizando os mais complexos, tornando-nos mais claros. A seguir é listada uma série de itens que complementam e/ou esclarecem o trabalho desenvolvido por Choy [14] e portanto são considerados como contribuição:

- Diagrama esquemático funcional de cada processo, detalhando cada operação envolvida;
- Apresentação das variáveis de entrada/saída em cada bloco de forma seqüencial;
- Apresentação dos processos utilizando-se as equações envolvidas;
- Diagrama esquemático geral do processo do filtro de análise LP;
- Apresentação dos procedimentos para a estimação do pitch;
- Detalhes do processo de extração das CW's;
- Algoritmo para a otimização do critério de alinhamento das CW's;
- Organização do envio de parâmetros do codificador para o decodificador;
- O resumo do estágio de análise;
- Descrição detalhada da geração das CW's e do pitch instantâneos;
- Diagrama esquemático geral do processo do filtro de síntese LP;
- O desenvolvimento do sistema de análise – síntese WI (padrão).

No capítulo de resultados, capítulo 5 deste trabalho, foi apresentado o *método da relação sinal ruído segmental com o auxílio da função de correlação cruzada normalizada* – método

SNRSEG-NCCF onde foi desenvolvido o algoritmo SNRSEG-NCCF para avaliar a reconstrução das formas de onda por segmento e para auxiliar na avaliação da defasagem entre os sinais originais e reconstruídos. No capítulo 5 também foi apresentada e avaliada a versão preliminar implementada do sistema de análise – síntese WI com interpolação das CW's nas posições regulares, ou *sistema de análise – síntese WI (com interpolação)*.

No apêndice A foi apresentada uma *proposta para melhoria na reconstrução do sinal* no estágio de síntese para: aprimorar o processo de localização e extração das formas de ondas características, as CW's; a partir da localização mais precisa das CW's, ajustar a estimação do pitch; e a partir das CW's localizadas com maior precisão e do pitch estimado ajustado, obter as CW's nas posições regulares pela interpolação. Parte desta proposta foi implementada no sistema de análise – síntese WI (com interpolação).

6.4 Trabalhos Futuros

Baseado nos estudos realizados, na seqüência são apresentadas algumas sugestões para continuidade da pesquisa:

I - Estratégias para melhorias no processo de localização e extração das CW's com ajuste no processo de estimação do pitch (proposta apresentada no Apêndice A)

Estas estratégias seriam aplicadas durante o estágio de análise (na camada externa), para aprimorar: **(i)** os processos de localização e de extração das formas de ondas características – CW's; **(ii)** o processo de estimação do pitch; e **(iii)** o processo de determinação das CW's nas posições regulares. O aprimoramento dos processos tem como objetivo uma codificação mais eficiente e, portanto visam uma melhor reconstrução do sinal residual e da síntese do sinal da fala. Em resumo as estratégias são:

- **(a)** Localização das CW's próximas aos pontos regulares de extração (localização com maior precisão);
- **(b)** Determinação e ajuste no valor do pitch das CW's localizadas (com maior precisão);
- **(c)** Determinação e correção no valor do pitch para as CW's a serem interpoladas nas posições regulares *(pitch por sub-quadro)*;
- **(d)** Determinação do valor do *pitch por quadro* (utilizando-se a regressão linear) a partir dos valores dos pitch das CW's localizadas e/ou dos valores do pitch ajustados para as CW's em posições regulares;
- **(e)** Determinação das *CW's nas posições regulares* ao longo do eixo de evolução das formas de ondas – eixo discreto no tempo (n), utilizando-se a interpolação linear.

II - Complementação do sistema WI

Completar o sistema WI com o estudo, a descrição e a visualização de um algoritmo para a segunda parte do sistema WI na camada interna do sistema WI, ou seja, realizar a compressão (codificação) e decodificação dos parâmetros. Integrar a camada interna com a camada externa completando o simulador para o sistema codificador/decodificador WI e partir para um aprimoramento geral do sistema WI. A melhoria geral seria principalmente no aprimoramento do algoritmo de estimação do pitch, no processo de interpolação das CW's e na quantização dos parâmetros.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Ramírez, M. A., and Minami, M., "*Low Bit Rate Speech Coding*". In: Proakis, J. G., (Ed.) Wiley Encyclopedia of Telecommunications. New York: Wiley, vol. 3, 2003, pp. 1299-1308.

[2] Kleijn, W. B., and Paliwal, K. K., "*An Introduction to Speech Coding*". In Kleijn, W. B., and Paliwal, K. K. Editors, Speech Coding and Synthesis, chapter 1, pp. 1-48. Elsevier Science B. V., Amsterdam, 1995.

[3] Jayant, N. S. and Noll, P., "*Digital Coding of Waveforms: Principles an Applications to Speech and Video"*. Englewood Cliffs, New Jersey: Prentice-Hall, 1984.

[4] Quackenbush, S. R., Bamwell, T. P., and Clements, M. A., "*Objective measure for speech quality"*. Englewood Cliffs, New Jersey: Prentice-Hall, 1998.

[5] ITU-T, "*Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs"*. ITU-T Recommendation P.862, February 2001.

[6] ITU-T, "*Objective Quality Measurement of Telephone-band (300-3400 Hz) Speech Codecs*". ITU-T Recommendation P.861, February 1998.

[7] A. H. et al, "*Articulation Testing Methods*", J. Acoust. Soc. Am., vol. 37, pp. 158-166, 1996.

[8] Kubichek, R., "*Standards and Technology Issues in Objective Voice Quality Assessment*", Digital Signal Processing: A Review Journal, vol. DSP, pp. 38-44, Apr. 1991.

[9] Wang, S., Sekey, A., and Gersho, A., "*An Objective Measure for Predicting Subjective Quality of Speech Coders*", IEEE J. Selected Areas in Comm., vol. 10, pp. 819-829, June 1992.

[10] Itakura, F. "*Line Spectrum Representation of Linear Predictive Coefficientes of Speech Signals*", Journal Acoustical Society America, vol. 57, p. S35, Apr. 1975. abstract.

[11] Moore, B. C. J., "*An Introdution to the Psychology of Hearing*". Academic Press, fourth ed., 1997.

[12] Cotanis, I., "*Speech Quality Evaluation for Mobile Networks*", Proc. IEEE Int. Conf. On Communications (Glasgow, UK), vol. 3, pp 1530-1534, May 2000.

[13] Beerends, J., and Stemmerdink, J., "*A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation*", J. Audio Eng. Soc., vol. 40, pp. 963-978, 1992.

[14] Choy, E. L. T., "*Waveform Interpolation Speech Coder at 4 kb/s*", Master of Engineering Thesis, Department of Electrical Engineering, McGill University, pp. 107, Montreal, Canada, Aug. 1998.

[15] Kleijn, W. B., and Haagen, J., "*Waveform Interpolation for Coding and Synthesis*". In *Speech Coding and Synthesis* (Kleijn, W. B., and Paliwal, K. K. editors), chapter 5, pp 175-207, Elsevier Science B. V., Amsterdam, 1995.

[16] Rabiner**,** Lawrence R. & Schafer, Ronald W. "*Digital Processing of Speech Signals"***,** New Jersey: Prentice-Hall, Inc., 1978. 512p.

[17] Deller**,** J. R; Proakis, J.G.,e Hansen, J.H. "*Discrete - Time Processing of Speech Signals"*, New York: Macmillan, 1993. 908p.

[18] Barnwell, T. P., Nayebi, K. and Richardson, C. H., "*Speech Coding: A Computer Laboratory Textbook*", (Barnwell, T. P., Hayes, M., Mersereau, R. M., and Smith, M. J. T editors), New York: John Wiley & Sons,1995, 184p.

[19] Makhoul, J; R, S. and Gish, H. "*Vector Quantization in Speech Coding*". Proceedings of IEEE, Vol. 73, n° 11, pp. 1551 - 1588, Nov. 1985.

[20] Soong, F., Juang, B., "*Line Spectrum Pair (LSP) and Speech Data Compression*", Proc. Of the Int. Conf. Acoust., Speech and Signal Processing (San Diego), pp. 1.10.1-1.10.4, Mar. 1984.

[21] Kabal, P. and Ramachandran, R. P., "*The Computation of Line Spectral Frequencies Using Chebyshev Polynomials*". IEEE Trans. ASSP, vol. 34, pp. 1419-1426, December 1986.

[22] Soong, F. K. and Juang, B., "*Optimal Quantization of LSP Parameters*", IEEE Trans. Speech and Audio Processing, vol. 1, pp. 15-24, Jan. 1993.

[23] Paliwal, K. K. and Kleijn, W. B., "*Quantization of LPC Parameters*". In *Speech Coding and Synthesis* (Kleijn, W. B., and Paliwal, K. K. editors), chapter 12, pp. 433-466. Elsevier Science B. V., Amsterdam, 1995.

[24] Hanzo L., Somerville, F. C. A. and Woodard, J. P., "*Voice Compression and Communications: Principles and Applications for Fixed and Wireless Channels*", (IEEE series on Digital & Mobile Communication, Anderson J. B., Series Editor). New York: John Wiley & Sons, 2001, 642p.

[25] Atal, B. S., Rabiner, L. R., "*Speech Research Directions*", AT&T Technical Journal, pp.75-87, 1986.

[26] Gersho, A., "*Advances in Speech and Audio Compression*", Proceedings of the IEEE. vol. 82, pp. 900-918, June 1994.

[27] Atal, B. S., Remde, J. R., "*A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates*", Proc. of the Int. Conf. Acoust., Speech and Signal Processing, pp. 6124-6127, 1982.

[28] Deprettere, E. F. and Kroon, P., "*Regular Excitation Reduction for Effective and Efficient LP-Coding of Speech*", Proc. of the Int. Conf. Acoust., Speech and Signal Processing, pp. 25.8.1-25.8.4, 1985.

[29] Kroon, P., Deprettere, E. F. and Sluyter, R. J., "*Regular-Pulse Excitation – A Novel Approach to Effective and Efficient Multi-Pulse Coding of Speech*", Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-34, nº 5, Aug. 1986.

[30] Atal, B. S., Schroeder, M. R., "*Stochastic Coding of Speech at Very Low Bit Rates*", Proc. of the Int. Conf. on Communication, pp. 1610-1613, 1984.

[31] Schroeder, M. R. and Atal, B. S., "*Code- Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates*", Proc. of the Int. Conf. Acoust., Speech and Signal Processing, pp.937-940, 1985.

[32] Spanias, A. S., "*Speech Coding: A Tutorial Review*", Proceedings of the IEEE. vol. 82, pp. 1541-1582, Oct. 1994.

[33] Jarvinen, K., Vainio, J., Kapanen, P., Honkanen, T., Haavisto, P., Salami, R., Laflamme, C., and Adoul, J. P., "*GSM Enhanced Full Rate Speech Codec*", Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing (Munich), vol. 2, pp. 771-774, 1997.

[34] Honkanen, T., Vainio, J., Jarvinen, Haavisto, Salami, R., and C.L. Adoul, J. P., "*Enhanced Full Rate Speech Codec for IS-136 Digital Cellular System*", Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing (Munich), vol. 2, pp. 731-734, 1997.

[35] DeJaco, A., Gardner, W., Jacobs, P., and LEE, C., "*QCELP: The North American CDMA Digital Cellular Variable Rate Speech Coding Standard*", Proc. IEEE Workshop on Speech Coding for Telecommunications (Ste. Adele), pp. 5-6, 1993.

[36] Atal, B. S., and Caspers, B. E., "*Beyond Multipulse and CELP Towards High Quality Speech at 4 kb/s*". In Atal, B. S., Cuperman, V., and Gersho, A., Editors, Advances in Speech Coding, pp. 191-201. Kluwer Academic Publishers, 1991.

[37] Tzeng, F. F., "*Analysis-by-Synthesis Linear Predictive Speech Coding at 4.8 kbits/s and Below*". In Atal, B. S., Cuperman, V., and Gersho, A., Editors, Advances in Speech Coding, pp. 135-143. Kluwer Academic Publishers, 1991.

[38] Stachurski, J., "*A Pitch Pulse Evolution Model for Linear Predictive Coding of Speech*", PhD Thesis, Department of Electrical Engineering, McGill University, pp. 140, Montreal, Canada, Feb. 1998.

[39] Goldberg, R. G., and Riek, L., "*Mixed Excitation Coding*", in *A* Practical Handbook of Speech Coders, chapter 11, pp. 157-191. CRC Press, Florida, 2000.

[40] McAulay, R. J., and Quatieri, T. F., "*Speech Analysis/Synthesis Based on a Sinusoidal Representation*", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-34, pp. 744-754, Aug. 1986.

[41] McAulay, R. J., and, Champion, T., "Improved Interoperable 2.4 kb/s LPC Using Sinusoidal Transform Coder Techniques". In Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing (Albuquerque, New Mexico), pp. 641-643, 1990.

[42] Griffin, D. W., "The Multi-band Excitation Vocoder", PhD Dissertation: MIT, Cambridge, MA, Feb. 1987.

[43] Griffin, D. W., and Lim, J. S., "A New Model-based Speech Analysis/Synthesis System". In Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing, pp. 513-516, Mar 1985.

[44] Griffin, D. W., and Lim, J. S., "A High Quality 9.6 kbps Speech Coding System". In Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing, 1986.

[45] Griffin, D. W., and Lim, J. S., "Multiband Excitation Vocoder". IEEE Trans. Acoust., Speech and Signal Processing, vol. 36, nº 8, pp. 1223-1235, Aug 1988.

[46] Brandstein, M. S., Monta, P. A., Hardwick, J. C., and Lim, J. S., "A Real Time Implementation of the Improved MBE Speech Coder". In Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing (Albuquerque, New Mexico), pp. 5-8, 1990.

[47] McCree, A., and Barnwell, T., "*A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding*". IEEE Trans. on Speech and Audio Processing, pp. 242-250, 1995.

[48] Supplee, L. M., Cohn, R. P., Collura, J. S., and McCree, A. V., "*MELP: The New Federal Standard at 2400 BPS*", Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing, pp. 1591-1594, 1997.

[49] McCree, A. V., Truong, K., George, E. B., Barnwell, T. P., and Viswanathan, V., "A 2.4 KBIT/S *MELP Coder Candidate for the New U. S. Federal Standard*", Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing, pp. 200-203, 1996.

[50] Kleijn, W. B. "*Continuos Representations in Linear Predictive Coding*". In Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc., pp. 201-204, Toronto, 1991.

[51] Kleijn, W. B., and Granzow, W., "*Methods for Waveform Interpolation in Speech Coding*". Digital Signal Processing, vol. 1, pp. 215-230. Jan. 1991.

[52] Shoham, Y., "*Low-rate Speech Coding Based on Time-frequency Interpolation*". In Proc. Int. Conf. on Spoken Language Processing, pp. 37-40, 1992.

[53] Shoham, Y., "*High-quality Speech Coding at 2.4 to 4.0 kbps Based on Time-Frequency Interpolation*", Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing (Minneapolis), pp. II167-II170, 1993.

[54] Kleijn, W. B., and Haagen, J., "*Transformation and Decomposition of the Speech Signal for Coding*". IEEE Signal Processing Letters, vol. 1, nº 9, pp. 136-138. Sept. 1994.

[55] Kleijn, W. B., and Haagen, J., "*Speech Coder Based on Decomposition of Characteristic Waveforms*". In Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing, pp. 508-511, Detroit, 1995.

[56] Kleijn, W. B., Shoham, Y., Sen, D., and Haagen, J., "*A Low Complexity Waveform Interpolation Coder*". In Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing, pp. 212-215, Atlanta, 1996.

[57] Tanaka, Y. and Kimura, H., "*A Low-bit-rate Speech Coding Using a Two-dimensional Transform of Residual Signals and Waveform Interpolation*". In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. I173-I176, Adelaide, 1994.

[58] Burnett, I. S. and Bradley, G. J., "*New Techniques for Multi-prototype Waveform Coding at 2.84 kb/s*". In Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing, pp. 261-264, Detroit, 1995.

[59] Jiang, Y. and Cuperman, V., "*Encoding Prototype Waveforms Using a Phase Codebook*". In Proc. IEEE Workshop on Speech Coding for Telecommunications (Annapolis), pp. 21-22, 1995.

[60] Festa, M. and Sereno, D., "*A Speech Coding Algorithm Based on Prototype Interpolation With Critical Bands and Phase Coding*". In Proc. European Conf. on Speech Commun. And Technology, pp. 229-232, Madrid, 1995.

[61] Tang, K. W. and Cheetham, B. M. G.,"*Fixed Bit-rate PWI Speech Coding With Variable Frame Length*". In Proc. IEEE Globecom Conf., pp. 1600-1603, 1995.

[62] Yang, G., Leich, H. and Boite, R. "Voiced *Speech Coding at Very Low Bit Rates Based on Forward-backward Waveform Prediction*". IEEE Trans. on Speech and Audio Processing, vol. 3, pp. 40-47, Jan. 1995.

[63] Sen, D. and Kleijn, W. B., "Synthesis Methods in Sinusoidal and *Waveform-Interpolation Coders*". In Proc. IEEE Workshop on Speech Coding for Telecommunications (Annapolis, MD), 1995.

[64] Shoham, Y. and Gersho, A. "*Pitch Synchronous Transform Coding of Speech at 9.6 kb/s Based on Vector Quantization*". In Proc. Int. Conf. Comm. (P. Dewilde and C. May, eds), (Amsterdam), pp. 1179-1182, 1984.

[65] Kubin, G., Atal, B. S. and Kleijn, W. B., "Performance of Noise Excitation for Unvoiced Speech". In Proc. IEEE Workshop on Speech Coding for Telecommunications (Sainte-Adele, Quebec), pp. 35-36, 1993.

[66] Parsons, T. W. "Voice and Speech Processing". New York: McGraw-Hill, 1987, 402p.

[67] Kleijn, W. B. "*Encoding Speech Using Prototype Waveforms*". IEEE Trans. on Speech and Audio Processing, vol. 1, pp. 386-399, Oct. 1993.

[68] Ramírez, M. A., "*A waveform Extractor for Scalable Speech Coding*". ". In Proc. IEEE of the Int. Conf. Acoust., Speech and Signal Processing, Hong Kong, vol. 2, pp. 169-172, 2003.

[69] Goldberg, R. G., and Riek, L., "*Speech Analysis Techniques*", in *A* Practical Handbook of Speech Coders, chapter 3, pp. 33-49. CRC Press, Florida, 2000.

Apêndice A

PROPOSTAS PARA MELHORIAS NO PROCESSO DE EXTRAÇÃO DAS CW'S NO SISTEMA DE ANÁLISE - SÍNTESE WI (PADRÃO)

A.1 Introdução

Neste apêndice são apresentadas as propostas que visam aprimorar o processo de localização e extração das CW's em um codificador WI convencional [14], para permitir que, em posições regulares ao longo do eixo de evolução das formas de ondas – eixo do tempo discreto (n), as CW's representem o sinal residual de forma mais confiável.

A.2 Propostas (P) – Na Análise: Durante a Preparação para a Extração das CW's

As propostas de inovações são aplicadas ao processo de extração das CW's durante o estágio de análise no codificador WI convencional descrito em [14]. As propostas têm como pontos básicos: **(a)** a precisão na localização das CW's próximas aos pontos regulares; **(b)** o ajuste no pitch destas CW's; **(c)** a correção no valor do pitch nas posições regulares utilizando-se a interpolação do pitch das CW's localizadas próximas ao ponto regular (por sub-quadro) ou utilizando-se o cálculo a partir da regressão linear sobre os pitch's das CW's localizadas (por quadro); **(d)** a determinação do pitch por quadro (utilizando-se a regressão linear) a partir dos pitch das CW's localizadas ou dos pitch's ajustados das CW's em posições regulares; **(e)** e a determinação das CW's nas posições regulares ao longo do eixo de evolução das formas de ondas – eixo do tempo discreto (n), (utilizando-se a interpolação linear dos coeficientes de Fourier - domínio da freqüência). Assim as propostas podem ser resumidas em **(P1)** e **(P2)**.

A.2.1 Proposta 1 (P1) – Ajustar o período do pitch (pitch por sub-quadro) durante a localização mais precisa das CW's para a extração. E ajustar o período do pitch por quadro após o ajuste do pitch por sub-quadro. Esta proposta envolve os itens **(a)**, **(b)**, **(c)** e **(d)**.

A.2.2 Proposta 2 (P2) – Interpolar, utilizando-se os coeficientes de Fourier da DTFS, as CW's nas posições regulares de extração ao longo do eixo de evolução das formas de ondas –

eixo do tempo discreto (n), após a localização mais precisa das CW's. Esta proposta envolve diretamente o item **(e)** mais os resultados obtidos dos itens anteriores, de **(a)** a **(d)**.

Observação: Na Figura A.21 são resumidas em um diagrama esquemático geral as operações e os procedimentos necessários para esclarecer e executar as propostas (P1) e (P2).

A.3 Motivação para as propostas (MP)

A motivação geral para as propostas baseia-se no fato de que o valor do pitch por quadro não é um valor exato considerando-se a dimensão do quadro em relação à dimensão do ciclo do pitch e que o período do pitch varia naturalmente a cada ciclo. A partir destas considerações surge a motivação para se determinar um valor mais preciso do pitch para localizar e extrair a CW, e ajustar o valor do pitch estimado por quadro.

Também as CW's são localizadas e extraídas em posições deslocadas das posições regulares, mas no lado do decodificador elas são consideradas em posições regulares, o que resulta em distorções do sinal na reconstrução. Assim surge também a motivação para interpolar as CW's nas posições regulares a partir das CW's localizadas, melhorando a representação do sinal residual. As motivações específicas para as propostas são descritas a seguir na motivação para a proposta 1 (MP1) e na motivação para a proposta 2 (MP2).

A.3.1 Motivação para a proposta 1 (MP1) – No processo de localização e extração das CW's em [14], o valor do pitch considerado (para localizar e extrair as CW's) são os valores interpolados linearmente a partir dos valores do pitch estimado por quadro. Mas em um quadro do sinal sonoro geralmente existem mais de um ciclo de pitch, onde cada um deles pode ter valor diferente do outro. Além disso, a variação no valor do pitch, ciclo a ciclo, pode não ser linear. Disso pode-se dizer que o processo de estimação do pitch por quadro não resulta em um valor exato para o pitch, e menos ainda para o valor interpolado por sub-quadro. Com base nessas premissas pode-se inferir que se durante a localização das CW's, próximas às posições regulares, o período do pitch for determinado com maior precisão, existirá também, como conseqüência, uma melhora na precisão para os valores de pitch nas posições regulares (utilizando-se a interpolação linear) e também para o valor do pitch por quadro (utilizando-se a linear usando o método dos mínimos quadrados).

A.3.2 Motivação para a proposta 2 (MP2) – No processo de localização de uma CW em [14], a posição de extração (ou a posição onde a CW está localizada) geralmente é deslocada

da posição regular que é obtida como posição para a extração da CW. Mas no lado do decodificador, considera-se que, a CW transmitida foi extraída em uma posição regular. Se as CW's forem interpoladas entre as CW's localizadas nas posições mais próximas nas respectivas posições regulares, pode-se então inferir que as CW's interpoladas representarão o sinal residual de forma mais confiável.

A.4 Objetivos das propostas (OP)

O objetivo geral das propostas (P1) e (P2) é permitir que as CW's representem o sinal residual com maior fidelidade nas posições regulares o que resultará em uma codificação mais eficiente e em uma melhor reconstrução do sinal no codificador. Assim os objetivos específicos das propostas são apresentados nos sub itens que seguem: objetivos da proposta 1 (OP1) e objetivos da proposta 2 (OP2).

A.4.1 Objetivos da proposta 1 (OP1) – Para a proposta (P1) os objetivos são: **(a)** Fazer uma localização com maior precisão da posição da CW próxima à posição (ou ponto) regular de extração; **(b)** Pesquisar um valor com maior precisão para o pitch correspondente à CW localizada; **(c)** E, a partir do ajuste no pitch para as CW's no quadro atual, realizar também a correção do pitch do quadro atual a ser transmitido ao decodificador.

A.4.2 Objetivos da proposta 2 (OP2) – Para a proposta (P2) o principal objetivo é obter as CW's que representam com maior fidelidade as CW's nas posições regulares de extração ao longo do eixo de evolução das formas de ondas – eixo discreto no tempo (n). A determinação de cada CW, na sua respectiva posição regular de extração, é realizada utilizando-se a interpolação linear entre as CW's localizadas consecutivamente (CW_anterior e CW_atual) em correspondência com as suas respectivas posições (reais) de extração. As CW's são representadas pelos coeficientes de Fourier da DTFS e portanto a interpolação é realizada no domínio da freqüência.

A.5 Considerações Iniciais Sobre o Pitch

Durante a análise, no codificador, o pitch é estimado por quadro e interpolado em intervalos regulares por sub-quadro, como está esquematizado nas Figuras A.1 e A.2.

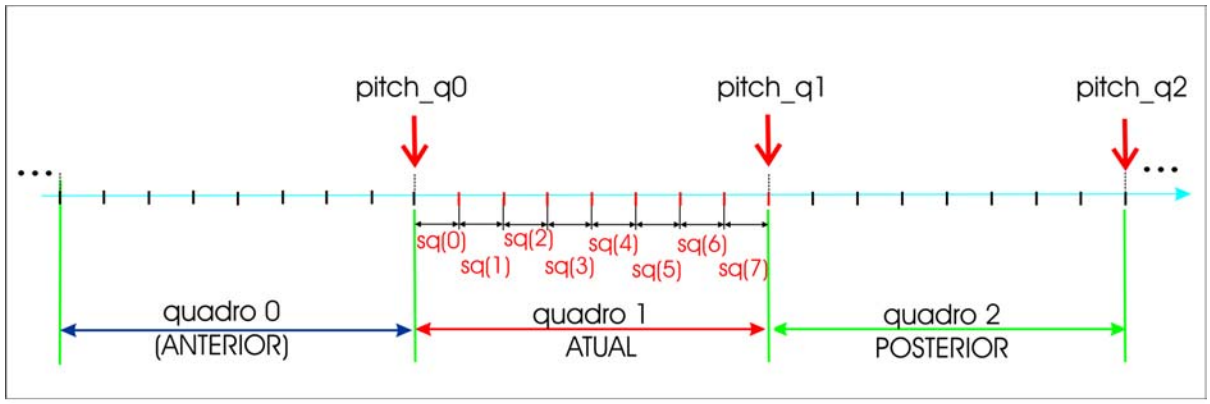

Figura A.1 – Diagrama esquemático representando o valor do pitch estimado por quadro. Cada valor é representado na posição da última amostra do quadro.

Desta forma o pitch no quadro atual é interpolado linearmente ao nível do sub-quadro como esquematizado na Figura A.2.

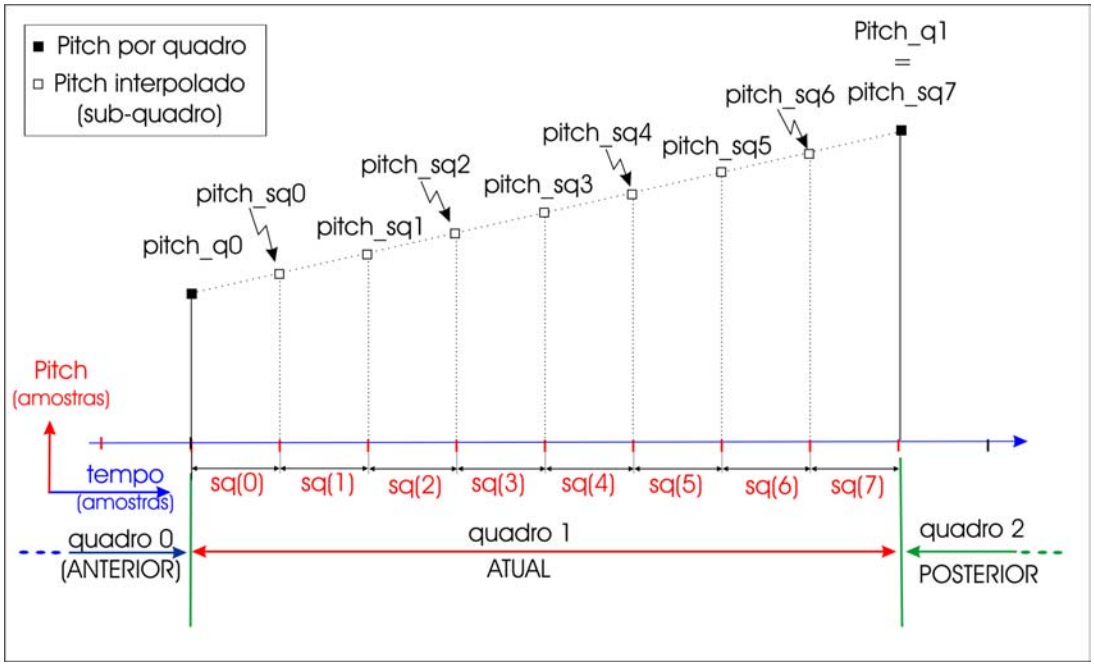

Figura A.2 – Diagrama esquemático representando a interpolação do pitch por sub-quadro, para o quadro atual, entre os valores do pitch do quadro anterior e do quadro atual. Os valores do pitch dos sub-quadros são posicionados na última amostra de cada sub-quadro.

A.6 Considerações Sobre a Localização das CW's

Nesta seção é feita uma breve descrição sobre o método de localização das CW's em [14] e o que se pretende acrescentar (ou inovar) no processo da localização das CW's para a extração.

A.6.1 Localização das CW's (Procedimentos em [14])

Durante a localização das CW's para a extração, usa-se uma janela J_{Ext} (Janela de Extração) com um comprimento fixo igual ao pitch_CW (pitch interpolado ao nível do sub-quadro) => $\{pitch\_sq0, pitch\_sq1, \cdots, pitch\_sq7\}$. Nos extremos da janela de extração (J_{Ext}) são fixadas as janelas para a verificação da energia, a janela J_{ee} no extremo esquerdo e a janela J_{ed} no extremo direito, cada uma delas com um comprimento $\delta_{máx} = 10$ amostras, como mostrado na Figura A.3. Estas janelas J_{ee} e J_{ed} são utilizadas para verificar a energia do sinal nas regiões dos limites de extração, e determinar a posição onde a soma da energia nas janelas seja mínima.

O centro (ou ponto médio) da janela J_{Ext} é considerado como seu ponto de referência para determinar sua posição nas amostras do sinal residual. Assim, ao conjunto $J_{Ext} + J_{ee} + J_{ed}$ é permitido um deslocamento de $-\varepsilon_{máx}$ até $+\varepsilon_{máx}$ ($\varepsilon_{máx} = 16$ amostras) onde o centro da janela J_{Ext} é posicionado inicialmente na *posição regular de extração*, a última amostra do sub-quadro.

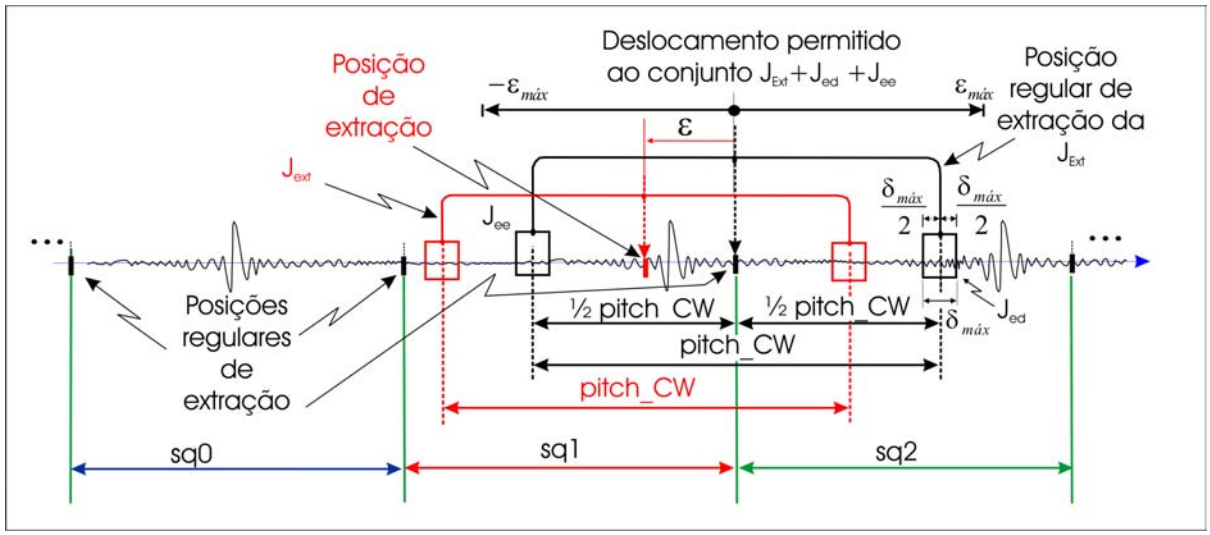

Figura A.3 – Diagrama esquemático – Mostra o processo de localização da CW em [14]. Na cor preta aparecem as janelas J_{Ext} (janela de extração), J_{ee} e J_{ed} (janelas de verificação da energia) colocadas na posição regular de extração, indicando o deslocamento permitido ao conjunto. Na cor vermelha observa-se a posição escolhida para a extração da CW de acordo com a posição das janelas J_{ee} e J_{ed} onde a soma da energia é mínima.

A.6.2 A Inovação na Localização das CW's

Em [14] a CW é localizada considerando-se o conjunto de janelas $J_{Ext} + J_{ee} + J_{ed}$ (interligadas). Ao conjunto de janelas é permitido deslocar onde o pitch _CW é considerado

constante, ou seja, a distância entre os centros de J_{ee} e J_{ed} é fixa e igual ao pitch_CW, como mostrado na Figura A.3.

Neste *novo modelo* usa-se os procedimento de [14] para fazer a *localização preliminar* da CW, como é mostrado na Figura A.3, e a partir dela, permite-se que as janelas J_{ee} e J_{ed} tenham deslocamentos independentes, dentro de uma certa faixa, na procura das regiões com energia mínima que indicarão os extremos da CW, fazendo assim a localização final (ajuste fino), como mostrado na Figura A.4.

A.7 A Localização das CW's

A partir das "Considerações sobre a localização das CW's", no item A.6, o processo de localização das CW's pode ser divido em duas fases:

- FASE 1- Localização preliminar da CW (procedimentos de [14]);
- FASE 2- Localização final da CW - Ajuste fino (A inovação).

A.7.1 FASE 1- Localização Preliminar da CW (Procedimentos de [14])

O objetivo nesta fase é localizar a CW_atual, correspondente ao sub-quadro atual, utilizando o respectivo pitch interpolado ao nível do sub-quadro.
Para realizar a localização são utilizadas as janelas J_{Ext} , J_{ee} e J_{ed}, e os procedimentos descritos em "Considerações sobre a localização das CW's", do item A.6 deste apêndice.
Realizar a localização da CW significa determinar a *posição da* J_{Ext} (ponto central da J_{Ext}) e os *limites da CW* (ponto médio de cada janela J_{ee} e J_{ed}), como mostrado na Figura A.3.

A.7.1.1 Características da Janela de Extração J_{Ext}

- *Função*: Localizar a posição e os limites da CW;
- *Tamanho*: Valor do pitch_CW (interpolado por sub-quadro);
- *Centro da janela*: Determina a posição da CW (após a localização) com 1/2 pitch à esquerda e 1/2 pitch à direita do ponto central da janela;
- *Deslocamento permitido*: De $-\varepsilon_{máx}$ até $+\varepsilon_{máx}$ ($\varepsilon_{máx} = 16$ amostras).

A.7.1.2 Características das Janelas de verificação da energia J_{ee} e J_{ed}

- *Função*: Localizar os extremos da CW utilizando-se o cálculo da energia do sinal limitado.

(Posição onde a soma da *energia do sinal nas duas janelas seja mínima*);

- *Tamanho*: $\delta_{máx} = 10$ amostras;
- *Posição*: As janelas J_{ee} e J_{ed} são ligadas pelos seus respectivos pontos médios aos extremos da janela J_{Ext}. Conseqüentemente cada janela J_{ee} e J_{ed} envolve $\delta_{máx}$ amostras com $\frac{\delta_{máx}}{2}$ amostras à esquerda e $\frac{\delta_{máx}}{2}$ amostras à direita do seu ponto central;
- *Deslocamento permitido*: Desloca-se as janelas interligadas à J_{Ext} e portanto com o mesmo deslocamento (de $-\varepsilon_{máx}$ até $+\varepsilon_{máx}$ ($\varepsilon_{máx} = 16$ amostras)).

Obs.: As janelas J_{ee} e J_{ed} mantêm uma distância fixa entre elas (entre os centros) do pitch_CW amostras.

A.7.1.3 O processo na FASE1

O processo na FASE1 é esquematizado nas figuras A.3 e A.4.

- Ao conjunto (J_{Ext} , J_{ee} e J_{ed}) é permitido deslocar na faixa de $-\varepsilon_{máx}$ até $+\varepsilon_{máx}$ ($\varepsilon_{máx} = 16$ amostras);
- Para cada posição calcula-se a energia do sinal limitado pelas janelas J_{ee} e J_{ed};
- A posição da CW (*posição preliminar –FASE 1)* é então determinada pela posição do conjunto (J_{Ext} , J_{ee} e J_{ed}) que tiver correspondência com a menor energia do sinal limitado pelas janelas J_{ee} e J_{ed};
- Assim a posição de extração da CW (*posição preliminar –FASE 1)*, o centro da J_{Ext}, e a posição (centro) das janelas J_{ee} e J_{ed}, ficam disponíveis para a próxima fase, FASE 2.

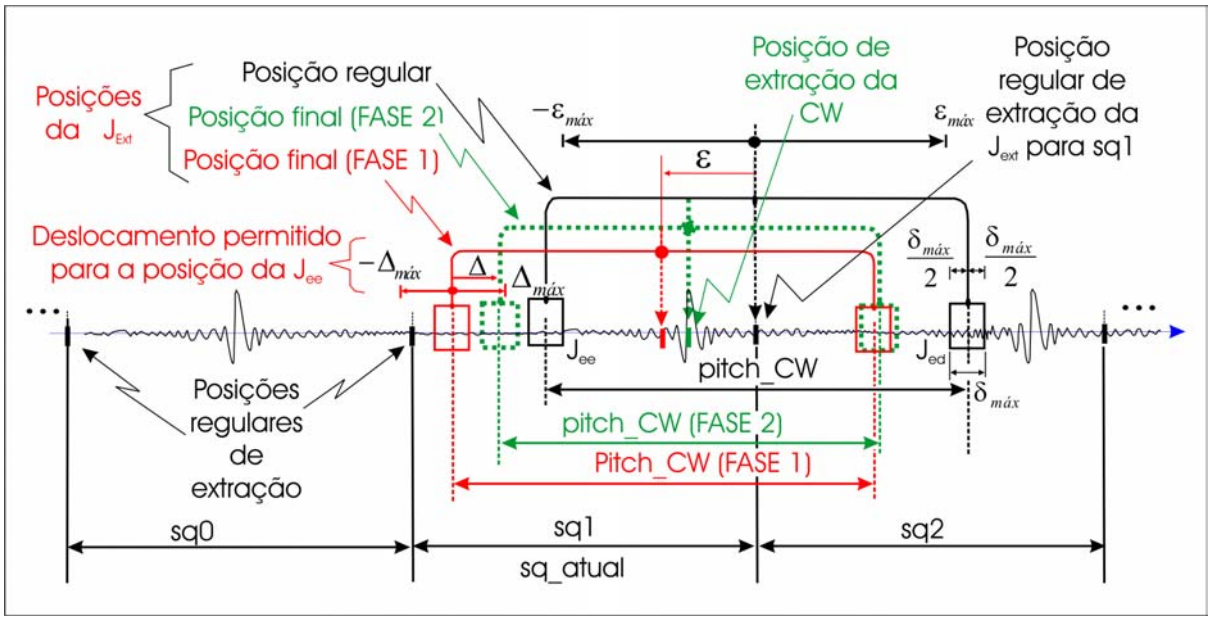

Figura A.4 – Diagrama esquemático dos processos de localização da CW- Processo em [14] relativo à FASE 1 e as inovações relativas à FASE 2. Na cor preta aparece o conjunto das janelas interligadas J_{Ext} , J_{ee} e J_{ed} colocado na posição regular de extração. Na cor vermelha aparece o conjunto das janelas colocado na posição final da FASE 1 correspondente ao processo em [14], onde pode-se observar a indicação do deslocamento independente, permitido para a janela J_{ee} a ser executado na próxima fase, a FASE 2 (o mesmo deslocamento independente, não indicado na figura, é permitido à janela J_{ed}). Na cor verde pode-se observar a posição final (FASE 2), para as janelas J_{ee} e J_{ed} nas posições onde a energia limitada é mínima para cada uma delas individualmente, e a janela resultante, a janela de extração final (FASE 2) que tem um comprimento de pitch_CW (FASE 2) medido entre o ponto médio de J_{ee} e de J_{ed}. A posição da CW no sinal residual é indicada como a posição média entre J_{ee} e J_{ed}.

A.7.2 FASE 2 - Localização Final da CW - Ajuste Fino (A Inovação)

O objetivo na FASE2 é fazer a localização da CW_atual com maior precisão, permitindo que as janelas para a verificação da energia, J_{ee} e J_{ed} possam ter deslocamentos independentes na procura individual dos extremos, onde a energia do sinal residual é mínima.

Na FASE2 o processo de localização começa com as janelas (J_{Ext} , J_{ee} e J_{ed}) posicionadas na *posição preliminar* determinada na FASE 1. O processo da FASE2 é esquematizado na Figura A.4. A cada janela J_{ee} e J_{ed} é permitido o deslocamento individual dentro da faixa de $-\Delta_{máx}$ até $+\Delta_{máx}$ ($\Delta_{máx} = 5$ amostras)[(A.1)] na determinação das posições onde sinal em cada janela tem energia mínima.

[(A.1)] A princípio foi adotado o valor de 5 amostras para $\Delta_{máx}$. Mas é necessário fazer alguns experimentos para verificar o impacto de $\Delta_{máx}$ na localização das CWs e nos limites admissíveis para a variação do pitch.

A flexibilização para os deslocamentos independentes das janelas J_{ee} e J_{ed}, pode resultar na variação do valor do pich_CW em uma faixa de pich_CW-$2\Delta_{máx}$ até o pich_CW+$2\Delta_{máx}$. Assim, a CW fica localizada a partir da determinação da posição de cada janela J_{ee} e J_{ed}, onde a energia é mínima. Os *extremos da CW*, que estão relacionados com as *posições das janelas* J_{ee} e J_{ed}, são determinados pelo *ponto médio* de cada uma das janelas (centro de cada janela). A distância entre os centros das janelas J_{ee} e J_{ed} determina o *comprimento da CW* que também corresponde ao *pitch_CW* da CW na FASE 2.

A posição média entre as janelas J_{ee} e J_{ed} indica a *posição da CW* ou posição da J_{Ext} no sinal residual.

A.7.2.1 O processo na FASE2

1- À cada janela J_{ee} e J_{ed},, é permitido o deslocamento individual, na faixa $-\Delta_{máx} \leq \Delta' \leq \Delta_{máx}$ em torno da posição do respectivo extremo da CW, determinados previamente na *posição preliminar* na FASE1.
2- Para cada posição Δ', calcula-se a energia limitada (E'_{limitada})[(A.2)].
3- A posição Δ que corresponde à posição da *janela de verificação da energia* com energia limitada mínima define-se a região onde se situa o extremo da CW. Calcula-se então o *ponto médio* (ou centro da janela) que define o extremo da CW. Aplica-se este procedimento à janela J_{ee} determinando o extremo esquerdo da CW e à janela J_{ed} determinando o extremo direito da CW.
4- O *pitch_CW* é determinado como a distância entre os extremos *(comprimento da CW)* .
5- Finalmente a posição média entre as janelas J_{ee} e J_{ed} (ou entre os extremos da CW) indica a *posição da CW* ou a posição da J_{Ext} no sinal residual.

A.8 Introdução aos Métodos de Preparação das CW's para a Codificação

As CW's com seus elementos (forma, pitch e posição) transportam a informação do sinal residual. Para a codificação visualiza-se duas situações que resultam em taxas de transmissão diferentes.

Na *primeira situação*, procura-se permanecer com a mesma taxa de atualização para o pitch utilizada em [14], transmitindo somente um valor do pitch por quadro. Para isto executa-se os procedimentos que ficam denominados de *Método I*, onde as CW's são localizadas,

---

[(A.2)] Nesta fase, definida como a energia do sinal limitado por uma das janelas de verificação de energia.

geralmente em posições deslocadas das posições regulares, e interpoladas nas posições regulares para serem codificadas e transmitidas ao decodificador.

Na *segunda situação* a ser considerada são procedimentos denominados de *Método II*, onde as CW's transmitidas ao decodificador são aquelas consideradas nas posições reais de extração onde foram localizadas. Para executa-lo é necessário transmitir a posição da CW, o pitch por CW e a CW propriamente (aqui podem surgir outras formas de preparar os parâmetros para a transmissão). Este método é mais preciso porém a taxa de transmissão é maior do que no Método I.

A.8.1 Método I – cálculo do pitch _CW interpolados nas posições regulares

No cálculo do pitch _CW interpolados nas posições regulares pode-se adotar um dos três procedimentos denominados de Procedimento I, Procedimento II ou Procedimento III.

A.8.1.1 Procedimento I

Calcula-se o pitch para a CW na posição regular (pitch_CW) utilizando-se a interpolação considerando-se os valores para o pitch_CW e a posição (da CW_anterior e da CW_atual) determinados na FASE 2 (no item A.7.2 deste apêndice). Duas das situações possíveis são esquematizadas nas Figuras A.5-a e A.5-b. Assim para cada duas CW's sucessivas, a CW_anterior e a CW_atual na FASE 2, determina-se por interpolação o pitch_CW(atual) na respectiva posição regular. O processo é executado até que o pitch seja interpolado em todas as posições regulares do quadro atual.

Para a interpolação do pitch na primeira posição regular no quadro atual utiliza-se também o pitch_CW e a pos_CW (FASE2) da última CW do quadro anterior.

A.8.1.1.1 O processo no Procedimento I

(a) Recebe as informações (pos_CW, pitch_CW) da última CW do quadro anterior localizada na FASE 2;
(b) Recebe as informações (pos_CW, pitch_CW) da CW localizada referente ao sub-quadro atual;
(c) Faz-se a interpolação (ou extrapolação) linear do pitch na posição regular correspondente ao sub-quadro atual;
(d) Repete-se os itens (b) e (c) até que o pitch seja interpolado (ou extrapolado) em todas as posições regulares do quadro atual;

(e) No final deste processo tornam-se disponíveis para o próximo processo (regressão linear), as informações do pitch_CW com as respectivas posições regulares (pos_reg).

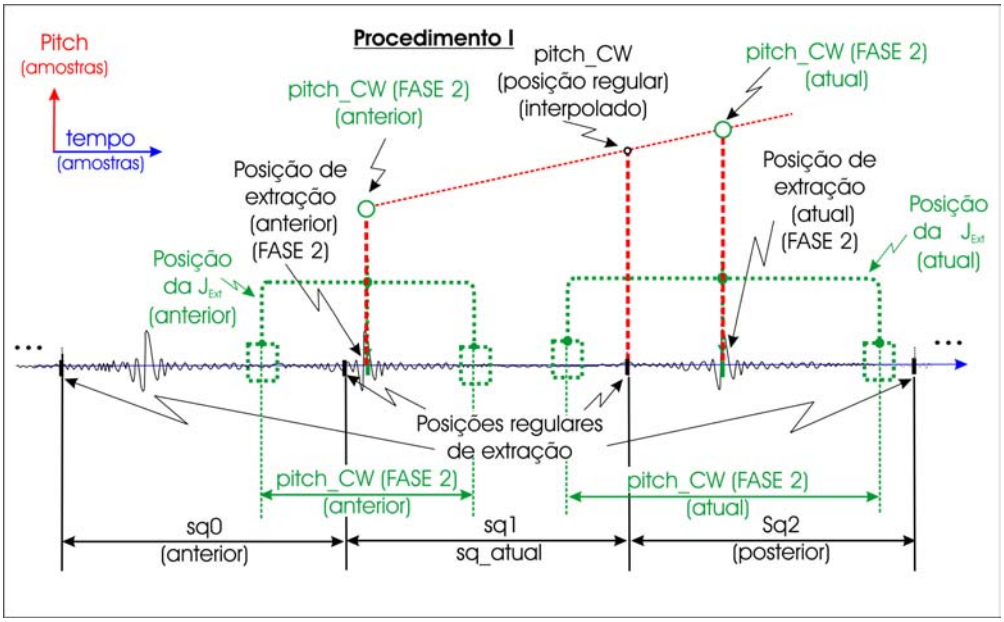

Figura A.5 (a) – Diagrama esquemático do Procedimento I para a interpolação do pitch_CW na posição regular a partir do pitch_CW e da respectiva posição, determinados na FASE 2, para duas CW's consecutivas, a CW_anterior e a CW_atual. Posição regular no intervalo entre as posições das CW's anterior e atual.

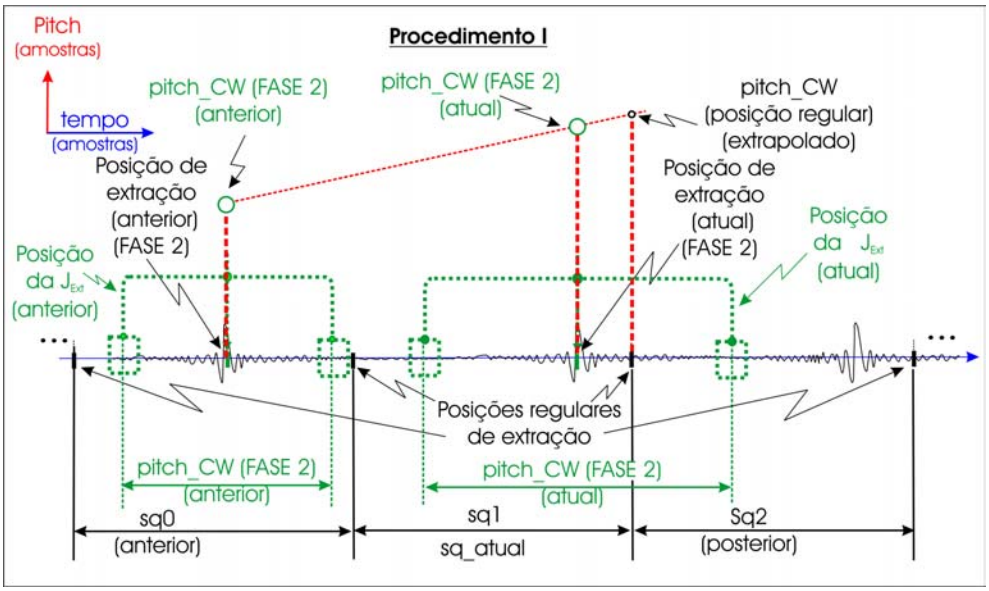

Figura A.5 (b) – Diagrama esquemático do Procedimento I para a interpolação (ou extrapolação) do pitch_CW na posição regular a partir do pitch_CW e da respectiva posição, determinados na FASE 2, para duas CW's consecutivas, a CW_anterior e a CW_atual. Posição regular fora do intervalo entre as posições das CW's anterior e atual, portanto ocorre extrapolação do pitch_CW.

A.8.1.2 Procedimento II

Executam-se os procedimentos da FASE 2 (explicados no item A.7.2 deste apêndice) localizando todas as CW's do quadro atual (determinando o pitch_CW e a posição da CW).
Faz-se a interpolação nas posições regulares para o pitch_CW, respeitando as posições das CW's localizadas que são adjacentes à respectiva posição regular, ou seja,

$$posição\_anterior(FASE2) \leq posição\_regular \leq posição\_posterior(FASE2).$$

Assim a posição regular está sempre entre as posições das CW's usadas na interpolação. Uma das situações possíveis é esquematizada na Figura A.6.
Para o processo de interpolação do pitch utiliza-se também o pitch_CW e a pos_CW (FASE2) da última CW do quadro anterior.

A.8.1.2.1 O processo no Procedimento II

(a) Recebe as informações (pos_CW, pitch_CW) da última CW do quadro anterior localizada na FASE 2;
(b) Recebe as informações (pos_CW, pitch_CW) de todas as CW's do quadro atual localizadas na FASE 2;
(c) Se a posição da última CW localizada no quadro atual for menor do que a última posição regular, é feita a localização da primeira CW no quadro posterior. Verifica-se novamente se a posição desta CW (no quadro posterior) é maior do que a última posição regular. Se for maior, segue-se com o processo de interpolação normal, se não, então é feita a extrapolação para a última posição. Desta forma considera-se as duas últimas CW's localizadas no quadro atual.
(d) Faz-se a interpolação linear do pitch em todas as posições regulares no quadro atual sempre considerando as CW's localizadas nas posições adjacentes à posição regular. A única exceção pode ocorrer na interpolação da última posição regular, como foi considerado no item (c) anterior;
(e) No final deste processo tornam-se disponíveis para o próximo processo (regressão linear), as informações pitch_CW com as respectivas posições regulares (pos_reg).

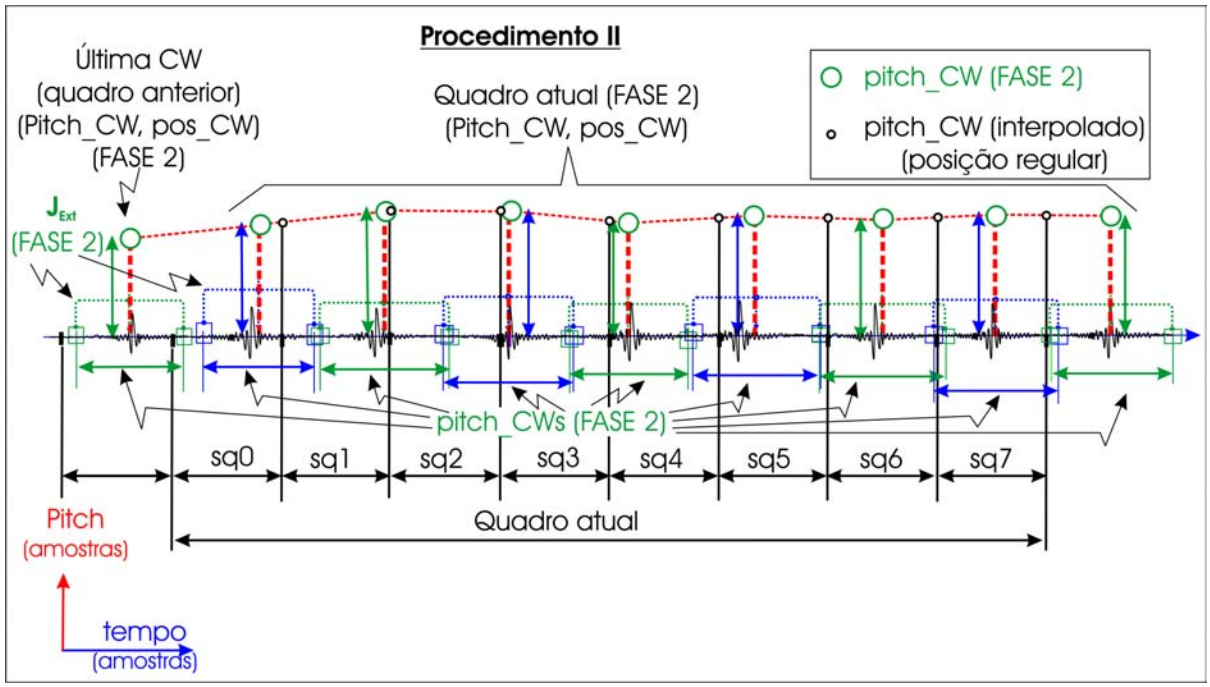

Figura A.6 – Diagrama esquemático do Procedimento II para a interpolação dos pitch_CW's nas posições regulares a partir dos pitch_CW's e das respectivas posições, determinadas na FASE 2.

A.8.1.3 Procedimento III

Executam-se os procedimentos da FASE 2 (explicados no item A.7.2 deste apêndice) localizando todas as CW's do quadro atual (determinando o pitch_CW e a posição da CW). Aplica-se então a regressão linear que pode ser realizada por um dos dois modos:

Modo I: Faz-se a regressão linear, método dos mínimos quadrados, a partir dos pitch_CW's e das posições das CW's (FASE 2), considerando também como ponto obrigatório da reta obtida o ponto para a posição e o pitch do quadro anterior [pos_q(anterior), pitch_q(anterior)].
Modo II: Faz-se a interpolação polinomial pela fórmula de Lagrange a partir dos (pitch_CW, pos_CW) FASE 2, na determinação dos pitch_interpolados nas posições regulares. Usando então os pitch_interpolados (nas posições regulares), os (pitch_CW, pos_CW) FASE 2 e também o ponto para a posição e o pitch do quadro anterior [pos_q(anterior), pitch_q(anterior)] determina-se a reta por regressão linear (método dos mínimos quadrados).
A partir da reta obtida, no Modo I ou no Modo II, calcula-se então os pitch_CW's linearizados nas posições regulares. Na Figura A.7 está esquematizado o Procedimento III no Modo I.

A.8.1.3.1 O processo no Procedimento III

(a) Recebe as informações:

(a1) o pitch do quadro anterior (pitch linearizado da última CW do quadro anterior) => pitch_q(anterior) em correspondência à posição regular do quadro anterior pos_q(anterior);

(a2) as posições e os pitch's (pos_CW, pitch_CW) de todas as CW's localizadas (FASE 2) no quadro atual.

(b) Aplica-se a regressão linear:

(b1) Modo I: É feita a regressão linear sobre todas as (pos_CW, pitch_CW) localizadas (FASE 2) considerando como ponto obrigatório da reta o ponto (pos_q(anterior), pitch_q(anterior));

(b2) Modo II: Aplica-se a fórmula de Lagrange sobre os pontos (pitch_CW, pos_CW) FASE 2 e determina-se os pontos (pitch_interpolado, pos_reg). A partir dos pontos (pitch_CW, pos_CW) FASE 2, dos pontos (pitch_interpolado, pos_reg) e do ponto (pos_q(anterior), pitch_q(anterior) determina-se a reta utilizando-se a regressão linear (métodos dos mínimos quadrados).

(c) Com a reta obtida, calcula-se os valores de pitch_CW (linearizados) nas posições regulares (Modo I ou Modo II).

(d) No final deste processo tornam-se disponíveis para o próximo processo (cálculo das CW's nas posições regulares) os valores de pitch_CW(linearizados) nas posições regulares (pos_reg), no Modo I ou Modo II, bem como o pitch_q (atual) que corresponde ao pitch_CW (linearizado) da última pos_reg do quadro atual.

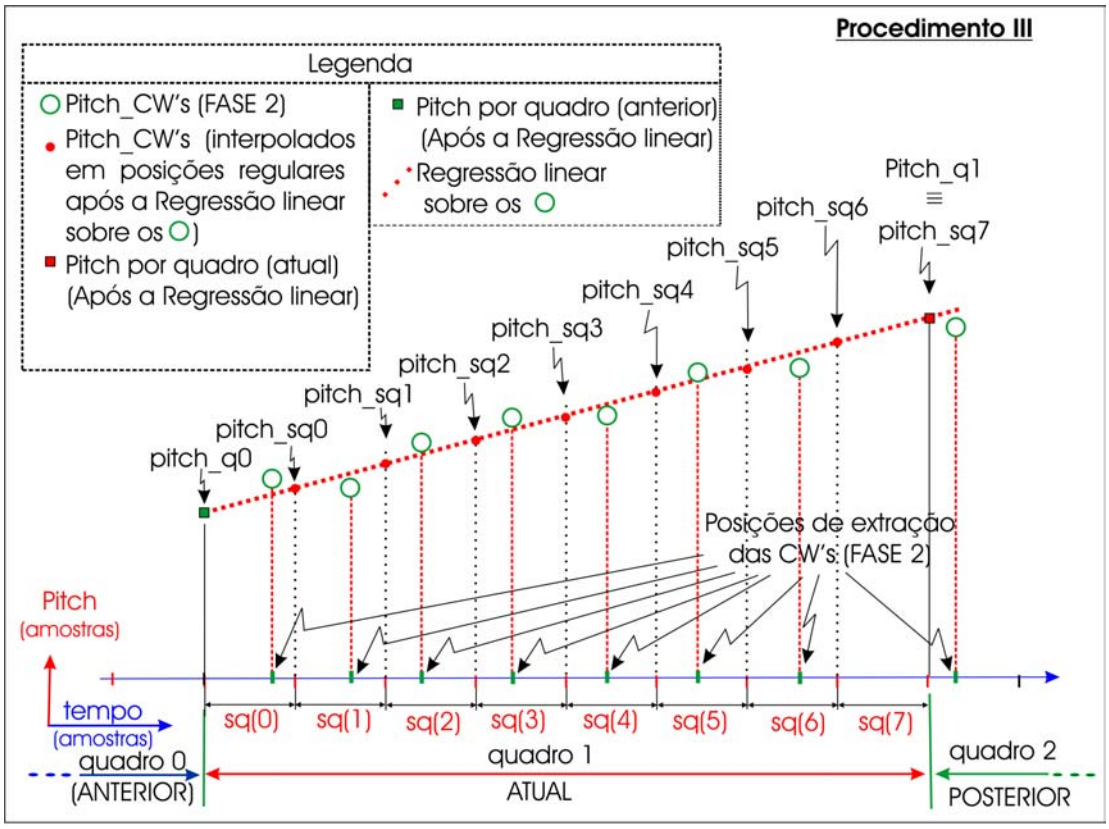

Figura A.7 – Diagrama esquemático do Procedimento III (Modo I). Mostra os dados (pitch_CW's, posição_CW's na FASE 2), o processo de regressão linear sobre os dados (pitch_CW's, posição_CW's) da FASE 2 e os valores calculados dos pitch_CW's_linearizados nas posições regulares a partir da regressão linear.

A.8.1.4 Observações sobre os procedimentos

Adotando o Procedimento I ou II, determina-se os valores dos pitch_CW's nas posições regulares.
Adotando o Procedimento III, determina-se os valores dos pitch_CW's já linearizados nas posições regulares.
A partir de experimentos será verificado qual dos procedimentos é o mais adequado.

A.8.2 Método II

Como foi descrito na introdução do item A.8 deste apêndice, os parâmetros para este método já estão prontos após o processo de localização das CW's FASE 2, onde ficam disponíveis as próprias CW's localizadas nas suas posições originais, as posições das CW's (pos_CW) e o pitch da CW (pitch_CW). Para executar este método é necessário transmitir a posição da CW,

o pitch por CW e a CW propriamente. Como foi comentado neste apêndice, podem surgir outras formas de preparar os parâmetros para a transmissão. Este método é mais preciso, porém a taxa de transmissão é maior do que no Método I. A intenção a princípio é implementar o método, transmitindo todos os parâmetros para fazer comparações com os resultados obtidos com o método I.

A.9 Regressão Linear dos pitch_CW's (Posição Regular) Relativo ao Método I

Após o cálculo do pitch_CW nas posições regulares (utilizando-se o Procedimento I ou o Procedimento II) aplica-se a regressão linear para determinar os valores dos pitch_CW's linearizados (nas posições regulares) e ajustar o valor do pitch por quadro a ser transmitido ao codificador. No caso do Procedimento III, mostrado na Figura A.7, os valores resultantes para os pitch_CW's são os pitch_CW's já linearizados (nas posições regulares). Portanto já estão prontos para o próximo processo (Cálculo das CW's nas posições regulares, explicados no item A.10 deste apêndice).

A.9.1 A regressão linear para os Procedimentos I e II

A partir dos valores dos pitch_CW's (interpolados nas posições regulares) determina-se os novos valores dos pitch_CW's na regressão linear (nas posições regulares) utilizando-se o método dos mínimos quadrados, ou seja os pitch_CW's (linearizados). Considera-se também como ponto obrigatório da reta obtida, o ponto relativo a posição e o pitch para o quadro anterior [pos_q(anterior), pitch_q(anterior)].

Na Figura A.8 está esquematizado o processo de regressão linear para os procedimentos I ou II.

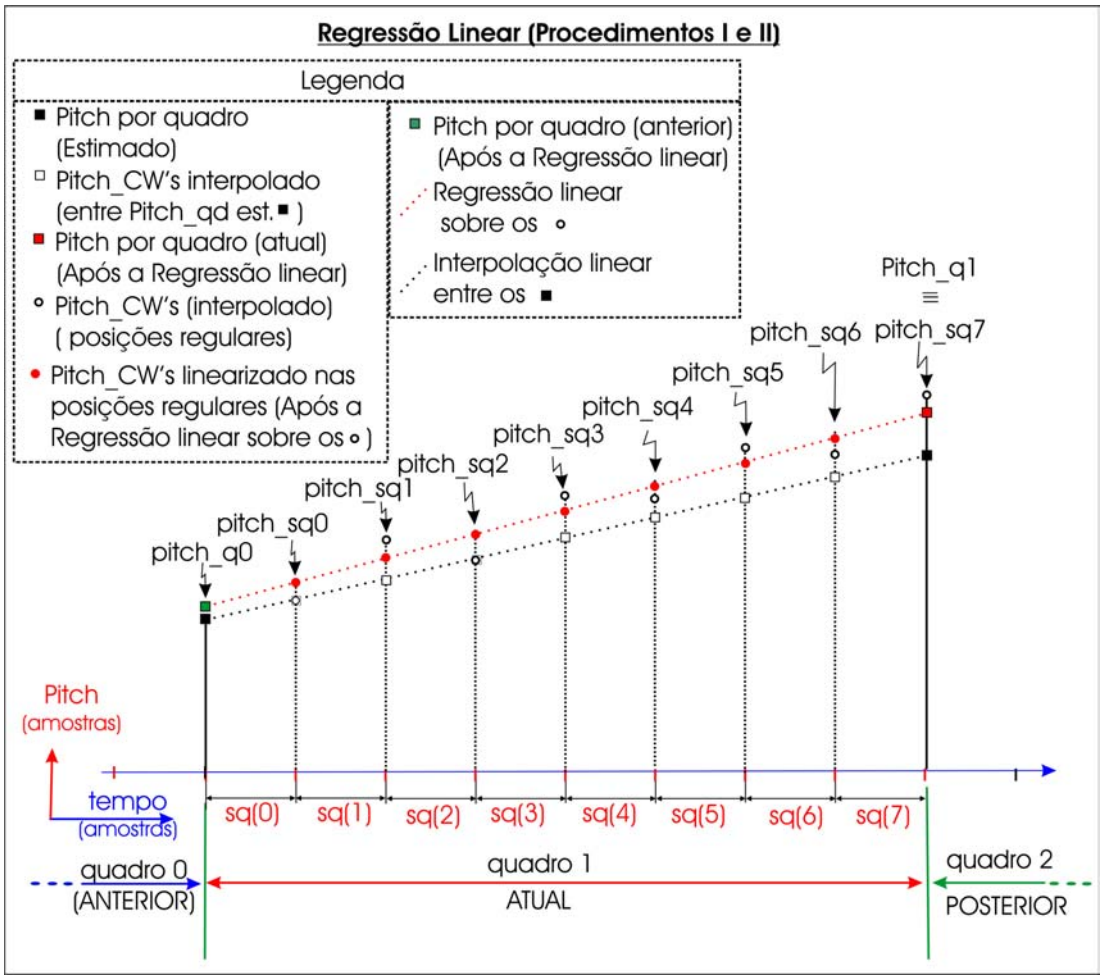

Figura A.8 – Diagrama esquemático. Mostra o processo de regressão linear aplicado aos valores de pitch_CW's, em posições regulares, calculados pelos Procedimentos I ou II.

A.9.1.1 O processo da regressão linear para os Procedimentos I e II

(a) Recebe as informações:

(a1) o pitch do quadro anterior (pitch linearizado da última CW do quadro anterior) => pitch_q(anterior) em correspondência à posição regular do quadro anterior pos_q(anterior);

(a2) os pitch (pitch_CW) interpolados nas posições regulares (pos_reg) do quadro atual;

(b) É feita a regressão linear sobre todos os (pos_reg, pitch_CW) interpolados considerando-se como ponto obrigatório da reta o ponto (pos_q(anterior), pitch_q(anterior));

(c) Com a reta obtida calcula-se os valores do pitch_CW (linearizados) nas posições regulares;

(e) No final deste processo tornam-se disponíveis para o próximo processo (cálculo das CW's nas posições regulares) os valores de pitch_CW (linearizados) nas posições regulares

(pos_reg), bem como o pitch_q (atual) que corresponde ao pitch_CW (linearizado) da última pos_reg do quadro atual.

Para os Procedimentos I, II ou III os valores dos pitch_CW's ficam então linearizados (pitch_CW's_linearizados) (ou alinhados em uma reta), o que permite a transmissão de um único valor de pitch (pitch_linearizado_quadro_atual) para o decodificador representando o pitch por quadro. Este valor é o pitch_CW_linearizado da última CW no quadro, relacionado com o último sub-quadro. No lado do decodificador os valores dos pitch_CW's são então recuperados utilizando-se a interpolação nas posições regulares entre os valores transmitidos do pitch_CW_linearizado (quadro anterior) e do pitch_CW_linearizado (quadro_atual).

A.10 Cálculo das CW's nas Posições Regulares (Relativo ao Método I)

Considerando a **CW_anterior** (pitch_CW, posição_CW e a CW_anterior[(A.3)]) e a **CW_atual** (pitch_CW, posição_CW e a CW_atual[(A.3)]) localizadas durante a FASE 2, mais o pitch_CW_linearizado (na posição regular) determinado com a regressão linear (no item A.9 para os Procedimentos I ou II ou no item A.8.1.3 para o Procedimento III), determina-se então a CW na posição regular (CW_pos_reg) utilizando-se a interpolação linear.

Durante o processo de interpolação as CW's (**CW_anterior** e **CW_atual**) são colocadas em suas posições reais de localização ao longo do eixo de evolução das formas de ondas – eixo discreto do tempo (n) e dispostas ao longo do eixo dos tempos m (para as amostras das CW's) perpendicular ao eixo (n), em formação para compor a superfície u(n,m) conforme mostrado na Figura A.9b. Com as CW's posicionadas (representadas por seus coeficientes de Fourier $\{A_k, B_k\}$ da "DTFS" – domínio da freqüência) executa-se a interpolação linear utilizando-se os respectivos coeficientes de Fourier, determinando-se então os coeficientes de Fourier relativos à CW interpolada na posição regular (CW_pos_reg).

Assim na determinação das CW's nas posições regulares são executadas as operações descritas na próxima subseção deste apêndice (processo da interpolação das CW'S) que estão esquematizadas nas Figuras A.9a, b, e c, e A.10-a e b.

A.10.1 Processo de Interpolação das CW's nas Posições Regulares (Relativo ao Método I)

(1) Recebe as informações:

[(A.3)] A CW_anterior e a CW_atual (a forma de onda) são representadas por seus coeficientes de Fourier da Série Discreta de Fourier no Tempo - Discrete Time Fourier Serie (DTFS).

(a) A **CW_anterior** (pitch_CW, posição_CW e a CW_anterior (Coeficientes da DTFS)). Inclusive a **CW_anterior** correspondente à última CW do quadro anterior localizada na FASE 2;

(b) A **CW_atual** (pitch_CW, posição_CW e a CW_atual (Coeficientes da DTFS));

(c) O pitch_CW_linearizado para a posição regular de interpolação.

(2) Considerando as **CW_anterior** e **CW_atual** alonga-se a CW mais curta (expansão temporal[A.4]), até que fique com o mesmo comprimento que a CW mais longa (utilizando-se a expansão espectral).

(3) Determina-se τ utilizando-se o critério de otimização do alinhamento, onde τ é o deslocamento (ao longo do eixo m) que a CW_atual deve ser submetida para que torne-se alinhada com a CW_anterior.

(4) Aplica-se τ na CW_atual, ou seja, faz-se o alinhamento da CW_atual com a CW_anterior.

(5) As CW's são colocadas em posição de formação da superfície u(n, m) conforme a Figura A.9b. Então executa-se a interpolação linear no domínio da freqüência, determinando a CW na posição regular (CW_pos_reg), ou seja, aplica-se a interpolação linear aos coeficientes de Fourier que representam as CW's atual e anterior, determinando os coeficientes de Fourier da CW interpolada.

(6) Expande-se[A.4] ou contrai-se[A.5] a CW na posição regular para um comprimento igual ao valor do pitch_CW_linearizado, de acordo com a comparação do valor do pitch_CW_linearizado na posição regular e o valor do pitch_CW da CW mais longa. Se o pitch_CW_linearizado for maior do que o pitch_CW da CW mais longa, então expande-se a CW interpolada na posição regular para o valor do pitch_CW_linearizado. Se ocorrer o inverso, contrai-se a CW interpolada na posição regular para o valor do pitch_CW_linearizado.

(7) No final deste processo tornam-se disponíveis os seguintes parâmetros para a compressão ou transmissão: as CW's nas posições regulares (CW_pos_reg), e o pitch_CW (quadro atual) (ou seja pitch_CW_linearizado da última CW do quadro atual).

[A.4] Como já foi descrito, no item 4.2.3.5 do capítulo 4 deste trabalho, Alinhamento das CW's, a **expansão espectral** de uma CW, que corresponde a uma **expansão temporal**, pode ser realizada por **(a) Inserção de zeros** entre os coeficientes DTFS para a expansão temporal periódica com um número inteiro de vezes; ou por **(b) acréscimo (ou apêndice) de zeros** à seqüência dos coeficientes de Fourier $\{A_k, B_k\}$ (após os coeficientes de maior ordem) para a expansão temporal de um comprimento menor do que um período.

[A.5] Para a **contração temporal**, trunca-se na seqüência $\{A_k, B_k\}$ os últimos coeficientes de Fourier (**truncamento espectral**), o que corresponde a eliminar as harmônicas de mais alta freqüência da CW. Ver também o item 4.2.3.5 do capítulo 4 deste trabalho, Alinhamento das CW's.

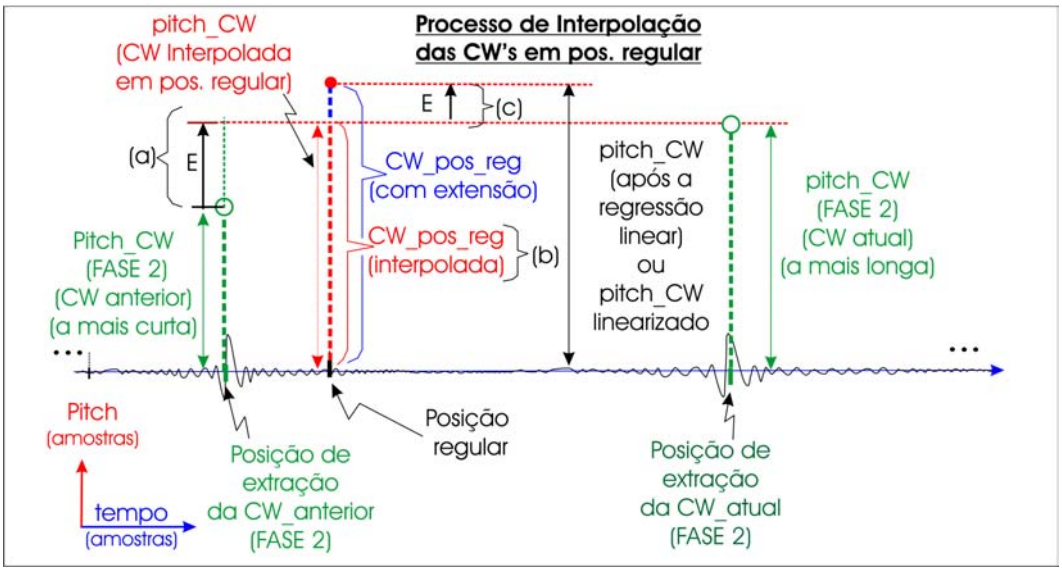

Figura A.9a – Diagrama esquemático para o processo de interpolação das CW's em posição regular. Neste caso a posição regular situa-se dentro do intervalo entre as posições das CW's anterior e atual. Aqui são indicadas as seguintes operações: (a) Expansão da CW mais curta; (b) Interpolação da CW na posição regular (CW_pos_reg) com o pitch_CW da CW mais longa; (c) Expansão (E) da CW_pos_reg para o valor do pitch_CW_linearizado.

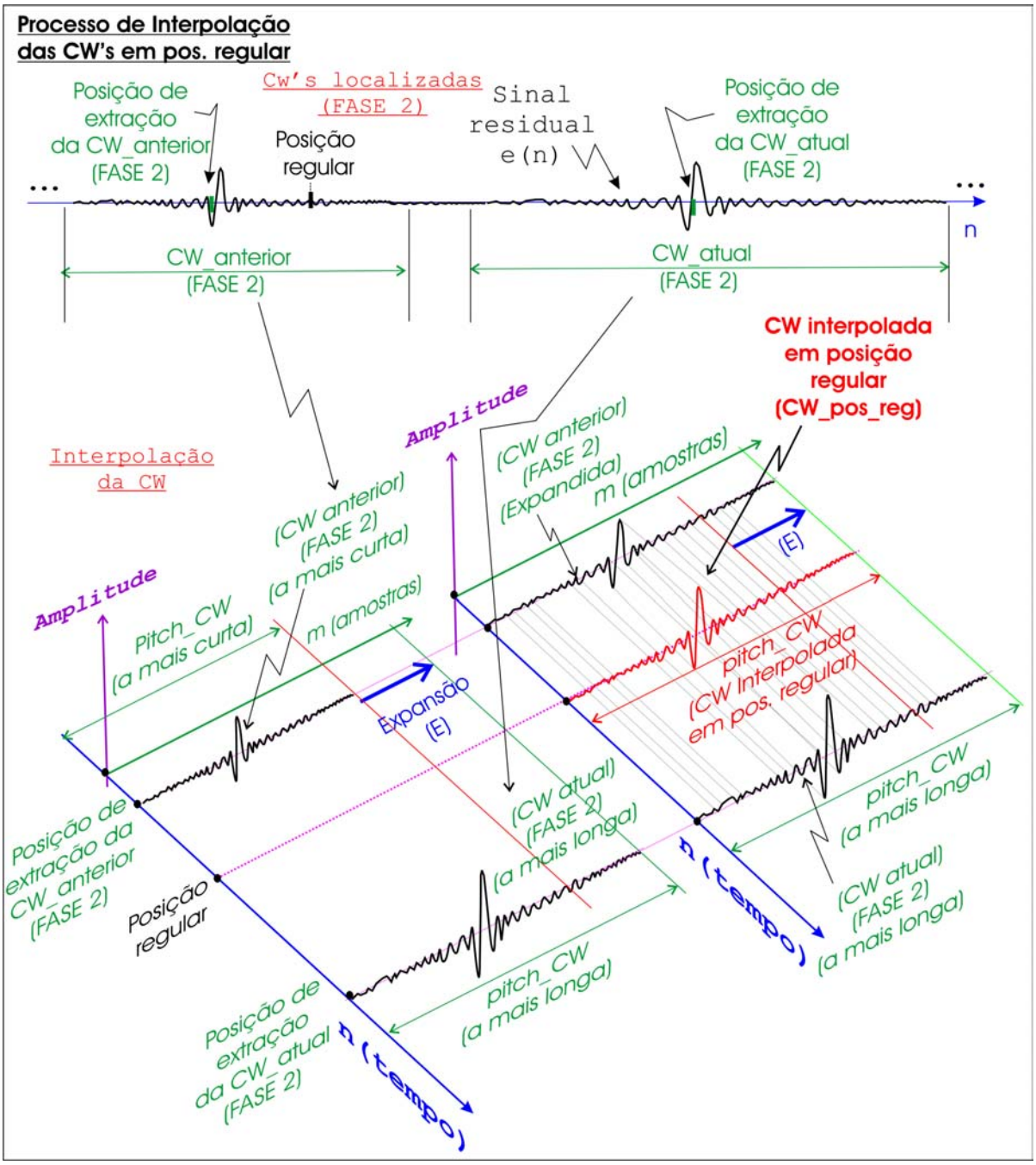

Figura A.9b – Diagrama esquemático para o processo de interpolação das CW's na posição regular. Mostra os detalhes correspondentes à Figura A.9a para as seguintes operações: (a) expansão da CW mais curta; e (b) Interpolação da CW na posição regular (CW_pos_reg) com o pitch_CW da CW mais longa.

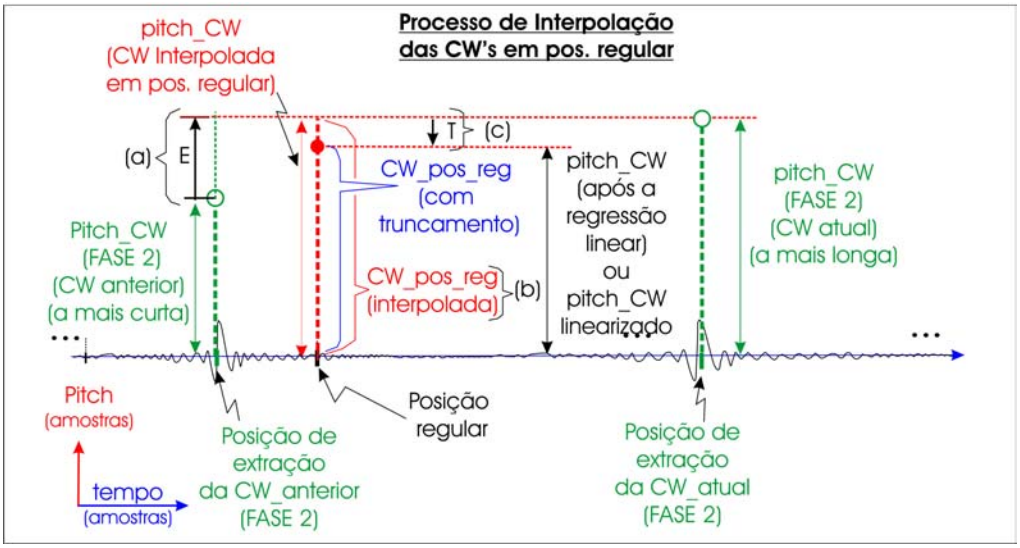

Figura A.9c – Diagrama esquemático para o processo de interpolação das CW's na posição regular. Neste caso a posição regular situa-se dentro do intervalo entre as posições das CW's anterior e atual. Aqui são indicadas as seguintes operações: (a) Expansão da CW mais curta; (b) Interpolação da CW na posição regular (CW_pos_reg) com o pitch_CW da CW mais longa; (c) Truncamento (T) da CW_pos_reg para o valor do pitch_CW_linearizado.

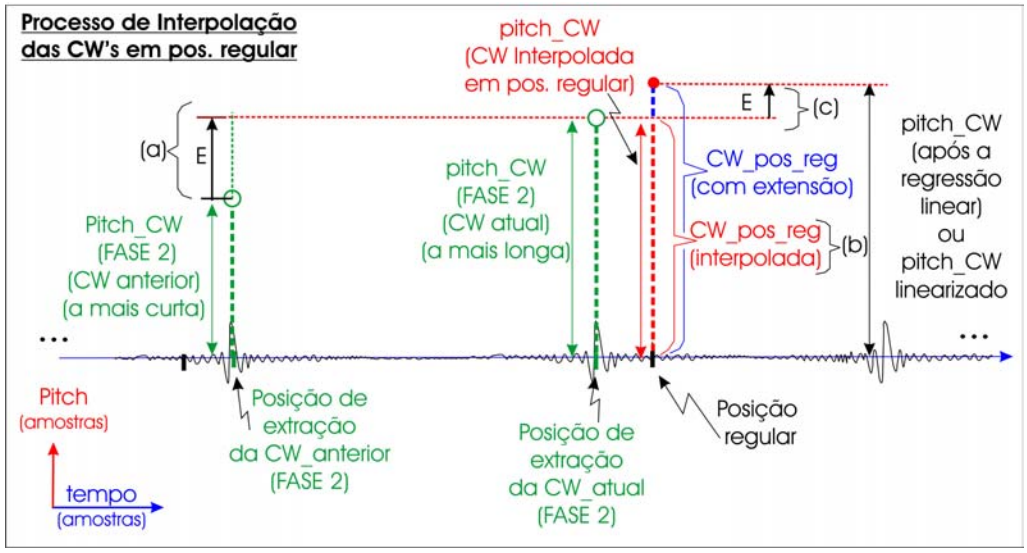

Figura A.10a – Diagrama esquemático para o processo de interpolação das CW's na posição regular. Neste caso a posição regular situa-se fora do intervalo entre as posições das CW's anterior e atual. Aqui são indicadas as seguintes operações: (a) Expansão da CW mais curta; (b) Interpolação da CW na posição regular (CW_pos_reg) com o pitch_CW da CW mais longa; (c) Expansão (E) da CW_pos_reg para o valor do pitch_CW_linearizado.

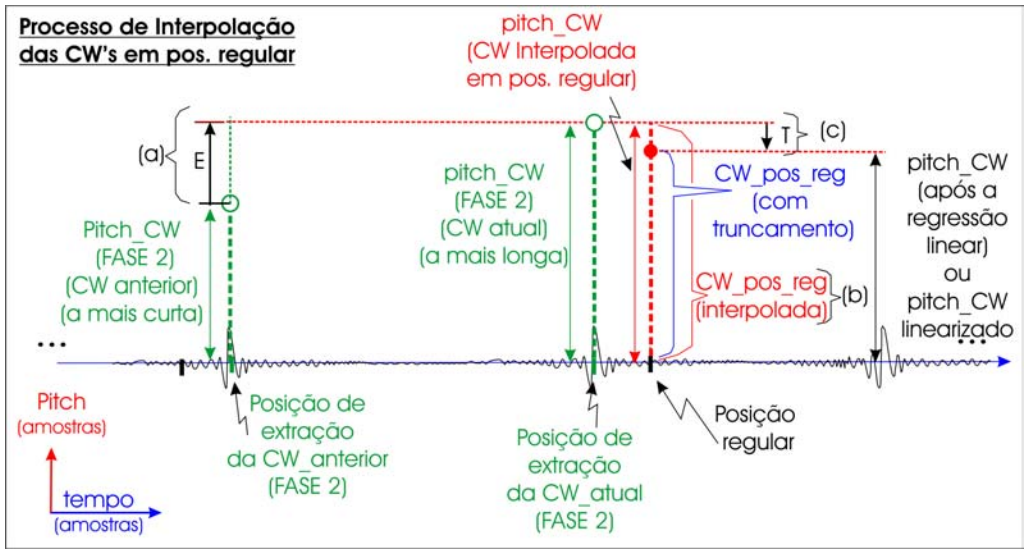

Figura A.10b – Diagrama esquemático para o processo de interpolação das CW's na posição regular. Neste caso a posição regular situa-se fora do intervalo entre as posições das CW's anterior e atual. Aqui são indicadas as seguintes operações: (a) Expansão da CW mais curta; (b) Interpolação da CW na posição regular (CW_pos_reg) com o pitch_CW da CW mais longa; (c) Truncamento (T) da CW_pos_reg para o valor do pitch_CW_linearizado.

A.11 Diagrama Esquemático Geral das Propostas P1 e P2

Aqui é mostrado o resumo em um diagrama geral para as seqüências de operações e procedimentos relativos às propostas P1 e P2.

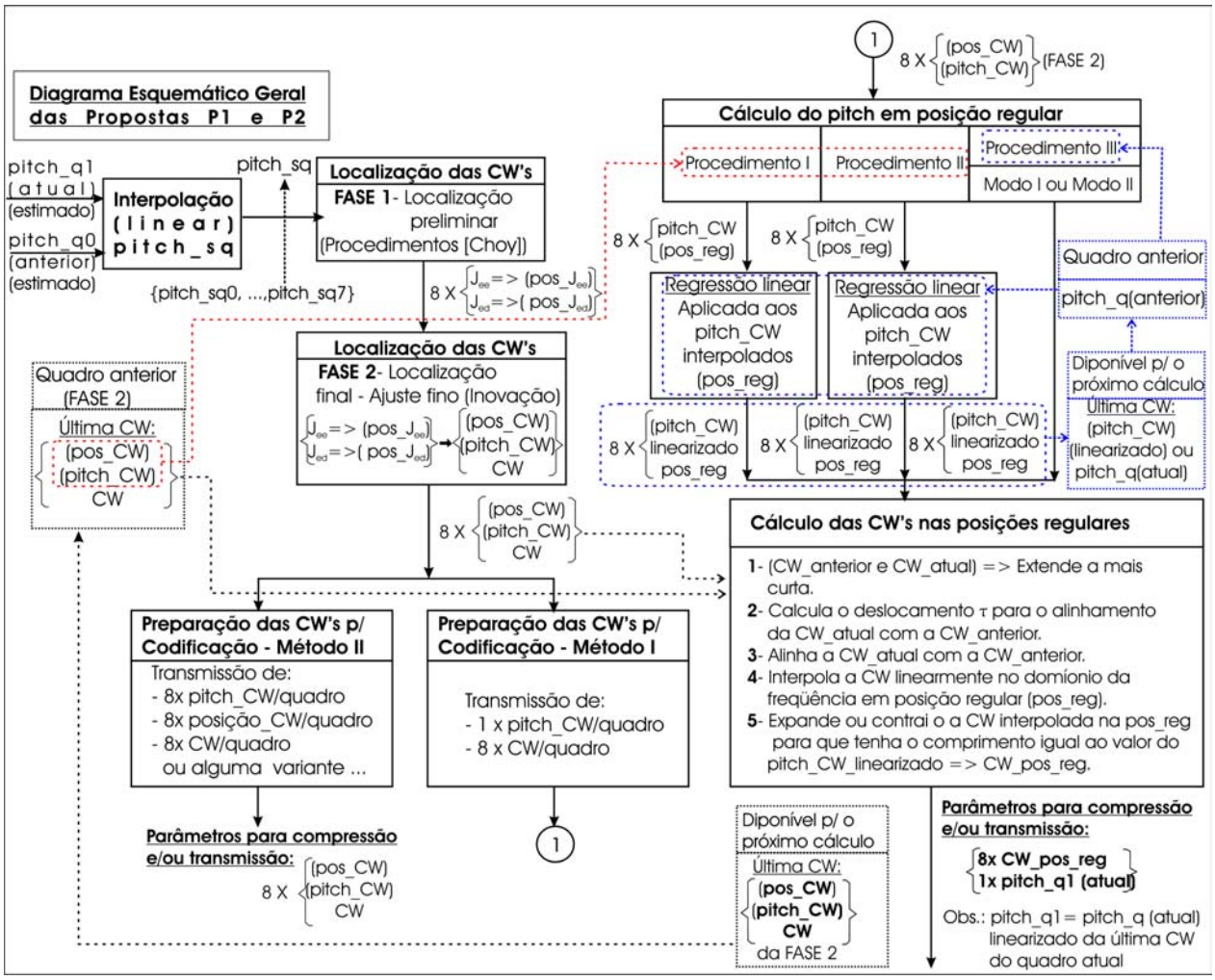

Figura A.11 – Diagrama esquemático geral das operações e dos procedimentos relativos às propostas P1 e P2.

A.12 Considerações finais deste apêndice

Neste apêndice foram apresentadas duas propostas, uma que visa aprimorar o processo de localização e extração das CW's utilizando-se a correção do pitch nas posições regulares e do pitch por quadro; e a outra que visa obter as CW's nas posições regulares ao longo do eixo de evolução das formas de ondas – eixo do tempo discreto (n) utilizando-se a interpolação, de forma que as CW's representem o sinal residual com maior fidelidade.

Na primeira proposta pretende-se ajustar o período do pitch durante a localização mais precisa das CW's para a extração. E ajustar o período do pitch por quadro após o ajuste do pitch por sub-quadro.

Na segunda proposta pretende-se interpolar as CW's nas posições regulares de extração (ao longo do eixo de evolução das formas de ondas) após a localização mais precisa para elas.

Para a localização mais precisa das CW's foi proposto o *Processo de Localização das CW's – FASE 2 (Ajuste fino)*. A eficácia deste processo é dependente do parâmetro $\Delta_{máx}$ que determina os limites de variação permitido para o pitch_CW. Devido a isso surge a necessidade de determinar que (quais) valor(es) $\Delta_{máx}$ pode assumir, através dos experimentos com uma gama representativa de sinais da fala, que resultem em uma localização mais precisa das CW's e conseqüentemente em valores mais precisos para o pitch_CW.

No processo da *Preparação das CW's para a Codificação* surgiram dois métodos, o Método I e o Método II. Nesse último método considera-se como parâmetros para a compressão e/ou transmissão as **CW's** (pos_CW, pitch_CW, CW (coeficientes da DTFS)) nas posições reais onde elas foram localizadas. A princípio este método será utilizado apenas como um padrão (como um modelo mais preciso – sem interpolação das CW's) para a comparação dos resultados a serem obtidos com o Método I, que considera a interpolação das CW's em posições regulares. No Método I os parâmetros finais para a compressão e/ou transmissão são as CW's (coeficientes da DTFS) – na taxa de 1 CW por sub-quadro, mais o valor do pitch_q (pitch do quadro atual). Neste método são apresentados três procedimentos Procedimento I, Procedimento II e o Procedimento III – (Modo I, Modo II) para o cálculo do pitch_CW nas posições regulares (pitch_CW_linearizado). Em todos eles utiliza-se o método de regressão linear – método dos mínimos quadrados, para a determinação do pitch_CW (ajustado-corrigido) nas posições regulares e o pitch_q (pitch do quadro atual – ajustado-corrigido) que é um parâmetro para a compressão e/ou a transmissão. O propósito é verificar (descobrir), a partir dos experimentos propostos na parte final destas considerações finais, qual dos Procedimentos é mais eficaz no processo final de reconstrução (ou síntese) do sinal da fala.

Após a determinação dos valores do pitch_CW_linearizado (nas posições regulares) e das **CW's** (pos_CW, pitch_CW, CW (coeficientes da DTFS)) localizadas (FASE 2) determina-se por interpolação linear dos coeficientes da DTFS as **CW's** (pos_reg, pitch_CW_linearizado, CW (coeficientes da DTFS)) no *Processo da Interpolação das CW's* (em posições regulares), que ao término da execução do Método I tornam-se disponíveis para a compressão e/ou a transmissão.

Para verificar (ou testar) a eficiência dos métodos propostos com o ajuste do pitch e com a interpolação das CW's nas posições regulares pretende-se utilizar alguns experimentos comparando os sinais da fala originais com os sintetizados e os sinais residuais no codificador

(original) com os sinais residuais no decodificador (reconstruído). Para o sinal da fala (original/sintetizado) serão aplicados: - o métodos de comparação visual do gráfico da forma de onda (amplitude x tempo); e - o método PESQ(A.7) [5] como uma medida objetiva da qualidade perceptual. Para o sinal residual (original-codificador/reconstruído-decodificador) serão aplicados: - os métodos de comparação visual do gráfico da forma de onda (amplitude x tempo); e - o método da SNRSEG (Relação Sinal Ruído Segmental) como medida objetiva da eficiência na reconstrução do sinal.

Portanto, pretende-se implementar as propostas incorporando-as ao codificador/decodificador WI convencional [14] realizando vários experimentos que visam comprovar e/ou testar os métodos (as hipóteses) e os procedimentos que foram descritos no desenvolvimento das propostas.

Vários testes (experimentos) poderão ser realizados, dentre eles pode-se citar:

(a) Determinar os limites admissíveis para a variação do pitch (pich_CW- $2\Delta_{máx}$ até pich_CW+ $2\Delta_{máx}$) variando-se $\Delta_{máx}$. Neste experimento serão utilizados alguns sinais da fala com variação nos tipos de expressões e dos locutores (masculino e feminino).

(b) Verificação/comparação da localização das CW's por inspeção visual no gráfico do sinal – amplitude x tempo para as CW's localizadas: (1) manualmente; (2) utilizando-se o codificador pelo novo método (relativo à FASE 2); e (3) utilizando-se o codificador pelo método convencional [14] (relativo à FASE 1). Neste experimento pretende-se mostrar com alguns exemplos a eficiência da localização das CW's utilizando-se o método proposto.

(c) Verificação/comparação dos valores do pitch:
(c1) pitch estimado por quadro e interpolado por sub-quadro (método convencional [14]);
(c2) pitch no Procedimento I;
(c3) pitch no Procedimento II;
(c4) pitch no Procedimento III (Modo I e Modo II);
(c5) cálculo manual do pitch pela inspeção visual no gráfico do sinal – amplitude x tempo.
Neste experimento pretende-se verificar qual dos *procedimentos* apresenta resultados mais próximos do cálculo manual do pitch.

(d) Verificação/comparação dos sinais da fala originais/sintetizados (Método I) pela visualização das formas de ondas(A.6) e também pela aplicação do método PESQ(A.7) [5]:

(d1) usando o Procedimento I;
(d2) usando o Procedimento II;
(d3) usando o Procedimento III (Modo I e Modo II);
(d4) e o método convencional [14].
Com o método PESQ pretende-se verificar qual dos Procedimentos é mais eficaz em relação ao sinal da fala sintetizado por uma avaliação objetiva da medida da qualidade perceptual da fala.

(e) Verificação/comparação dos sinais residuais no codificador/decodificador (reconstruído) (MétodoI) pela visualização das formas de ondas[(A.6)] e também pela SNRSEG (Relação Sinal Ruído Segmental):
(e1) usando o Procedimento I;
(e2) usando o Procedimento II;
(e3) usando o Procedimento III (Modo I e Modo II);
(e4) e o método convencional [14].
Com a SNRSEG pretende-se verificar qual dos Procedimentos é mais eficiente na reconstrução do sinal residual reconstruído.

(f) Verificação/comparação dos sinais da fala originais/sintetizados (Método II) pela visualização das formas de ondas[(A.6)] e também pela aplicação do método PESQ[(A.7)] [5]. Pretende-se avaliar o sinal da fala sintetizado, usando o Método II, pela avaliação objetiva da medida da qualidade perceptual da fala.
(g) Verificação/comparação dos sinais residuais codificador/decodificador (reconstruído) (MétodoII) pela visualização das formas de ondas[(A.6)] e também pela SNRSEG (Relação Sinal Ruído Segmental). Pretende-se avaliar a reconstrução do sinal residual reconstruído utilizando-se o Método II.

(h) Verificação/comparação dos sinais da fala sintetizados pela visualização das formas de ondas[(A.6)] e também pela aplicação do método PESQ[(A.7)] [5] para medir a qualidade perceptual. Comparar os resultados obtidos entre o Método I (item d) e o Método II (item f):
(h1) Método I (procedimento I – item d1) e Método II (item f);
(h2) Método I (procedimento II – item d2) e Método II (item f);

[(A.6)] Comparação e visualização dos sinais pelos gráficos da amplitude x tempo.
[(A.7)] O método Perceptual Evaluation of Speech Quality (PESQ) que foi adotado em 02/2001 como recomendação P.862 do International Telecomunication Union (ITU-T) [5] em substituição à recomendação P.861 que utiliza o método Perceptual Speech Quality Measure (PSQM) [6]. O PESQ foi adotado como método objetivo para fazer a predição da medida perceptual da qualidade da fala. Ele é baseado na representação psicoacústica dos sons da fala.

(h3) Método I (procedimento III (Modo I, Modo II) – item d3) e Método II (item f).

(i) Verificação/comparação dos sinais residuais no decodificador pela visualização das formas de ondas e também pela da Relação Sinal Ruído Segmental (SNRSEG). Comparar os resultados obtidos entre o Método I (item e) e o Método II (item g):
(i1) Método I (procedimento I– item e1) e Método II (item g);
(i2) Método I (procedimento II– item e2) e Método II (item g);
(i3) Método I (procedimento III (Modo I, Modo II) – item e3) e Método II (item g).

Observação: Para aplicação do método SNRSEG pretende-se verificar se existe defasagem entre o sinal residual (no codificador) e o sinal residual no decodificador (reconstruído) por inspeção visual do gráfico dos sinais (amplitude x tempo) e utilizando-se método da correlação cruzada normalizada entre os sinais. Caso seja constatada a defasagem ela deve ser corrigida para todo o sinal ou por segmento conforme for verificado se a defasagem é constante ou não para os segmentos.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo