UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
ESCOLA POLITÉCNICA
DEPARTAMENTO DE ELETRÔNICA E DE COMPUTAÇÃO

Marcação Automática de Eventos em Transmissões de TV

Autor:
Luiz Gabriel Lins Bentes Mendonça de Vasconcelos

Orientador:
Prof. Sergio Lima Netto, Ph.D.

Examinador:
Prof. Eduardo Antonio Barros da Silva, Ph.D.

Examinador:
Vagner Luis Latsch, M.Sc.

DEL
Março de 2008

Agradecimentos

Meus sinceros agradecimentos:

- ao Professor Sergio Lima Netto, pela orientação dada durante todo período de desenvolvimento deste projeto;
- ao Professor Luiz Wagner Pereira Biscainho, pela sugestão de utilizar o *pitch* neste projeto;
- à minha família, pelo apoio incondicional;
- à minha namorada, pelo apoio e compreensão nos momentos de ausência;
- aos meus colegas de turma, pela amizade e ajuda durante esses anos de graduação;
- à Central Globo de Engenharia da TV Globo e seus funcionários, pelo material fornecido e toda ajuda necessária ao desenvolvimento deste projeto;
- a todos funcionários e professores da Universidade Federal do Rio de Janeiro, por todos os serviços prestados.

Resumo

O objetivo deste projeto é desenvolver um método capaz de encontrar os melhores momentos de um programa esportivo, especificamente futebol, através do áudio do narrador. A proposta é identificar na narração do locutor esportivo os momentos em que a energia e a freqüência fundamental da sua voz aumentam suficientemente para caracterizarem bons momentos. Para tal, sucessivas análises serão realizadas para definição do método. Além disto, será implementado um aplicativo com interface gráfica que aplique todo o estudo realizado e teste-o com o restante da base de dados que não foi utilizada para o desenvolvimento. O estudo realizado neste projeto é interessante para emissoras de TV que têm diversos recursos consumidos com o processo melhores momentos em transmissões televisivas.

Palavras-chave: Identificação de Melhores Momentos, *Pitch* e Energia de Voz, Transmissão de Futebol.

Abstract

The objective of this project is to develop a method that finds highlights of a sport program, specifically soccer transmissions, through the audio of the narrator. The proposal is to identify in the speaker's narration the moments where the energy and the pitch frequency of the voice signal increase significantly enough to characterize the desired highlight. For such a task, successive analyses are performed for a complete definition of the method's setup. Moreover, an application with graphical interface was developed implementing the algorithms previously determined, allowing one to test the tool with the remaining database (other games, other narrators etc.) not used during the tool's development. The final system is interesting for broadcasting companies that have a lot of resources consumed in processing the highlight marking and identification.

Key-words: Highlight identification, Voice Energy and Pitch, TV Broadcast.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1 - Introdução

No mundo de hoje, e a cada dia que passa, maior é a necessidade humana por consumo das mais variadas maneiras de entretenimento, tais como Internet, shopping centers, peças teatrais, cinematográficas, prática de esportes e viagens. A mídia, além da missão de informar, também tem o objetivo de entreter a população. No Brasil, altos índices de audiência são alcançados em programas esportivos. Sabendo disso, as emissoras de TV não poupam tempo, recursos e esforços para atender a essa parcela da população, o que resulta em um consumo enorme de tais fatores. Assim, se torna altamente atrativo a utilização de tecnologias para tentar otimizar a produção de programas esportivos. O interesse em programas esportivos é tão relevante, que, hoje em dia, diversas transmissões acontecem simultaneamente e programas secundários os acompanham exibindo e noticiando acontecimentos que ocorreram durante a transmissão do programa, tais como, no caso de uma partida de futebol, gols, faltas, pênaltis, cartões etc.

Atualmente, especificamente para selecionar eventos de programas, há em cada transmissão um operador acompanhando e marcando os eventos em um sistema. Levando-se em conta que várias partidas acontecem simultaneamente, o número de operadores necessários para a sinalização de bons momentos cresce consideravelmente. Seguindo a linha da economia de esforços, recursos e tempo, seria extremamente útil o desenvolvimento de tecnologias que pudessem facilitar a marcação dos bons momentos. Uma tecnologia que conseguisse automatizar esse processo seria de grande valia, reduzindo os esforços consideravelmente.

De ínicio, sempre se pensa em utilizar processamento de imagens para tentar extrair informações de uma transmissão televisiva. Porém, é notório que obter informações tão específicas de uma imagem, tais como bons momentos de um evento esportivo, seria extremamente difícil. Além do fato de cada programa esportivo ter características visuais próprias do que seria um bom momento. Portanto, devido à especificidade de cada programa e dificuldade de generalização do problema por meio do processamento de imagens, surge o áudio como alternativa à solução do problema. É sabido que em toda transmissão há um narrador, e que sempre que há um bom momento, esse narrador aplica emoção em sua voz, independentemente do tipo de evento esportivo que está sendo transmitido. Para identificar especificamente o bom momento, mesmo pelo áudio, seria necessário reconhecer as palavras ditas pelo narrador, mas reconhecer que houve um bom momento é um objetivo atingível e que pode ser útil para uma etapa posterior de marcação. Neste projeto, após considerar viável extrair as informações desejadas através do processamento de áudio, foi estudado e desenvolvido um método e um sistema para marcar os bons momentos de partidas de futebol.

1.1 Proposta de trabalho

Este Projeto Final propõe o estudo e desenvolvimento de um método que consiga marcar automaticamente bons momentos de uma partida de futebol. Para começar é necessário estudar maneiras de extrair as informações desejadas da voz do narrador. Analisando a Figura 1, é possível ver que o modelo de produção da voz

humana pode ser representado pela excitação dos pulmões, pelo período de *pitch* das cordas vocais que podem estar vibrando ou não e pelo filtro que modela o trato vocal. Neste modelo, é preciso saber quais desses parâmetros usar para descobrir se o narrador se exaltou ou não. A excitação certamente é uma das responsáveis pelas alterações na voz quando ocorre um bom momento, já que ela é diretamente responsável pelo volume da voz. O período de pitch relaciona-se à entonação da voz, então nela também é possível ver alterações ocorrendo um bom momento. Já o filtro do trato vocal contém informações mais presas ao conteúdo linguístico da voz, sofrendo poucas variações em um bom momento.

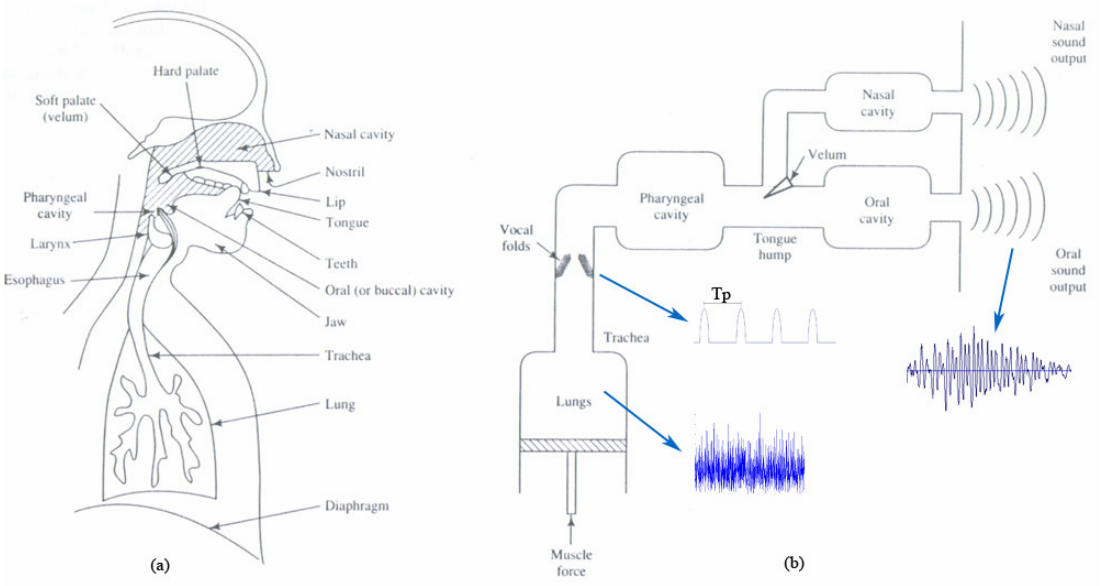

Figura 1 - Orgãos e elementos do corpo humano responsáveis pela geração da voz, representado bilogicamente, e (b) representado em blocos. [1]

Assim, dois sinais de informação serão trabalhados, onde um contém a variação da energia ao longo do tempo e o outro a variação do *pitch* no tempo. Assim, tendo em mãos os sinais de informação, o próximo passo é desenvolver um módulo de decisão de bom momento ou não. Por fim, após desenvolver um método de seleção de bons momentos a partir de um sinal de áudio, tem-se de validar o sistema com diversos materiais oriundos do mesmo narrador e de outros narradores. A partir do resultado, será visto se o sistema deve ser automático ou semi-automático, o que tornaria o método um pouco menos seleto e deixando a cargo de um operador descartar os trechos selecionados erroneamente, mas mesmo neste caso a tarefa seria muito mais simples, podendo ser feita em muito menos tempo.

1.2 Base de Dados

O método proposto neste projeto será construído realizando análises em sinais reais de vídeo que contenham programas esportivos, mais precisamente partidas de futebol. Algumas transmissões foram cedidas pela TV Globo, e estão listadas na Tabela 2.Os três narradores listados na Tabela 1 farão parte do estudo.

Tabela 1 - Narradores

Narrador I	Eduardo Moreno
Narrador II	Galvão Bueno
Narrador III	Cléber Machado

Tabela 2 – Jogos que compõem a base de dados.

	Partida	**Narrador**
Sinal I	Vasco x Flamengo 1° Tempo	Narrador I
Sinal II	Vasco x Flamengo 2° Tempo	Narrador I
Sinal III	Chivas Guadalajara x San Jose	Narrador I
Sinal IV	Botafogo x Vasco	Narrador II
Sinal V	Brasil x Chile	Narrador II
Sinal VI	Boca Jrs. X Grêmio	Narrador III

As transmissões são sinais digitais de vídeo com áudio *embedded*. Os arquivos contêm um *stream* de áudio digital amostrado à taxa de 48.000 Hz com 16 bits por amostra em dois canais. O canal esquerdo referente ao sinal da narração, e o direito ao ambiente. Utilizar somente o canal com o sinal da narração diminui o efeito dos ruídos que o ambiente poderia ocasionar na narração. Porém, o microfone do narrador também se encontra no ambiente, assim, os ruídos do ambiente se somam à voz do narrador, apesar de serem bem reduzidos.

Por fim, na Tabela 3, vemos a quantidade de bons momentos existentes em cada partida.

Tabela 3 - Bons Momentos em cada partida, em números absolutos.

Sinal	Narrador	Bons Momentos
Sinal I	Narrador I	14
Sinal II	Narrador I	15
Sinal III	Narrador I	20
Sinal IV	Narrador II	28
Sinal V	Narrador II	9
Sinal VI	Narrador III	6

1.3 Organização do Texto

O Capítulo 2 estuda mais a fundo a energia do sinal de voz com o propósito de conseguir gerar um sinal de informação que seja capaz de dizer os momentos considerados bons pela energia. Testes são feitos para avaliar como a energia deve ser medida ao longo do tempo para permitir uma análise subseqüente adequada.

O Capítulo 3, assim como o Capítulo 2, estuda a fundo métodos de detecção de *pitch*, informação esta que será usada para avaliar a característica de bom momento ou não. São necessárias análises para verificar a influência de trechos com silêncio e não sonoros na formação do sinal de informação.

O Capítulo 4 explica o módulo de decisão do sistema, que usa os sinais de informações obtidos nos Capítulos 2 e 3 para decidir entre bom momento ou não. Considerações sobre a classificação de acordo com a proximidade de dois trechos ou mesmo com a duração de um trecho são feitos para permitir uma avaliação mais robusta.

O Capítulo 5 apresenta a ferramenta desenvolvida e realiza testes com o método, aplicando os materiais da base de dados para validar o método e o sistema. Nesse capítulo é tomada a decisão de tornar o sistema automático ou semi-automático. Em uma primeira etapa será validado o jogo que serviu de base para o desenvolvimento do sistema, para depois validar jogos que tenham o mesmo narrador, e por fim validar jogos com diversos narradores de características diferentes.

O Capítulo 6 mostra um resumo de toda a dissertação, com comentários a respeito dos resultados obtidos, contribuições do projeto e propostas para trabalhos futuros.

Capítulo 2 - Sinal de Informação Baseado na Energia

Determinar em que trechos de um sinal de voz encontram-se bons momentos através da energia, a princípio, parece uma tarefa simples. Porém, por se tratar de uma aplicação particular dos fundamentos de análise de sinais, em que diversos métodos e variações terão de ser ponderados, a tarefa torna-se intensa. Neste capítulo, após uma formulação geral do problema, são revistos métodos capazes de gerar sinais auxiliares que contenham informações características de bons momentos. Estes métodos serão adaptados para a aplicação particular deste trabalho. Neste contexto, descrevem-se as implementações dos métodos formulados, seguido de análises que avaliarão cada método e definirão inicialmente os parâmetros necessários para a implementação do sistema.

Na seção 2.1 serão analisadas, de forma geral, as diversas maneiras de gerar um sinal contendo informações que caracterizam os bons momentos a partir da energia de um sinal de narração real, passando por Energia e Janela, Deslocamento da Janela e Escalamento. A seção 2.2 aponta, a partir dos resultados obtidos com o estudo da seção 2.1, quais métodos e parâmetros são os mais indicados para a classificação de bom momento ou não. Por fim, a seção 2.3 resumirá o capítulo e discutirá os resultados obtidos.

2.1 Formulação do Problema

2.1.1 Energia e Janela

A princípio, quando pensamos em analisar energia, calculamos a energia instantânea do sinal utilizando a equação (2.1) apresentada em [2]. Apesar de simples, esse modo de gerar o sinal de informação baseado na energia não nos dará dados substanciais para determinar onde há ou não bom momento, pois o sinal de energia praticamente acompanha o sinal de voz. Mesmo sabendo disso, o método de energia instantânea será estudado nesse capítulo.

$$E(n) = x^2(n) \tag{2.1}$$

Outra maneira (não tão instantânea) é considerar amostras que estão ao seu redor. Para isso, extraem-se os segmentos do sinal usando janelas de comprimento *N*, como ilustrado na Figura 2, e calculando a energia *E* para cada janela da forma indicada na equação (2.2).

$$E = \sum_{n=1}^{N} x^2(n) \tag{2.2}$$

O objetivo do janelamento é encontrar períodos que tenham informação de bom momento. Assim, temos que observar as características dos sinais de áudio em bons momentos para podermos determinar o tamanho *N* adequado do segmento.

O valor de *N* é um importante parâmetro de nossa análise. Um valor pequeno gera um número exagerado de segmentos, o que torna o processamento subseqüente

muito pesado computacionalmente. Um valor grande, por outro lado, dificultaria a determinação exata do início e do fim de um momento de interesse.

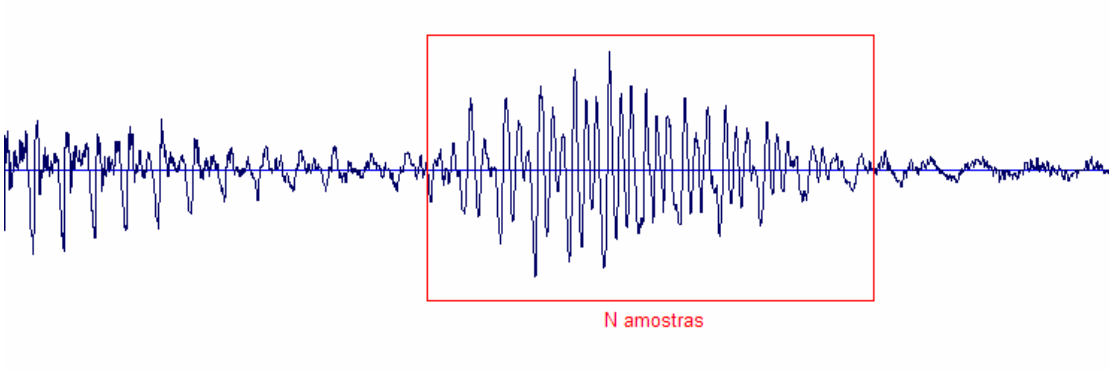

Figura 2 - Exemplo de uma Janela Retangular de *N* amostras em um sinal de voz qualquer.

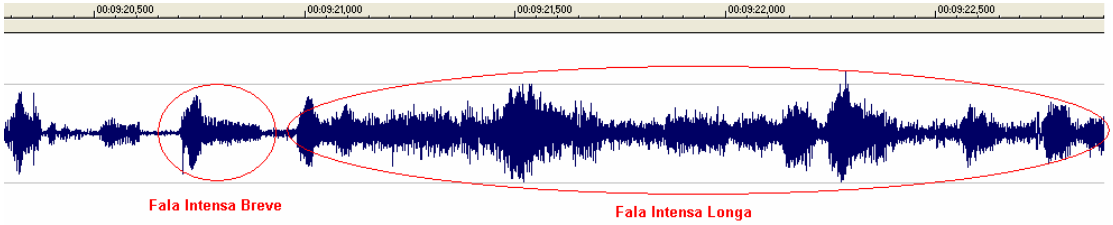

Figura 3 - Exemplo de duração de cerca de 200ms de uma fala intensa breve e cerca de 2s de um uma fala intensa longa.

2.1.2 Deslocamento da Janela

Após o cálculo da energia da janela, temos de deslocá-la para percorrer o sinal inteiro. Há duas maneiras de fazer isso, deslocamento superposto e não superposto no tempo, representadas na Figura 4. Em geral, a segunda forma pode ser vista como um caso particular da primeira, onde o deslocamento *M* é igual (ou maior) ao tamanho da janela *N*.

Com o deslocamento superposto o sistema tem a vantagem de ser mais preciso na escolha de um ponto a ser marcado. Porém, claramente vemos que o tempo de processamento dele é maior que o do deslocamento não superposto. Por exemplo, para um sinal de cinqüenta amostras, com uma janela de tamanho $N = 10$ amostras, são necessários cinco deslocamentos de janela não superposta $M = N$ para percorrer o sinal inteiro, enquanto que com o deslocamento superposto de uma em uma amostra $M = 1$ são necessários cinqüenta deslocamentos.

Entretanto, o tempo de processamento no janelamento superposto pode ser facilmente otimizado utilizando o algoritmo *buffer* circular. Ele se inicia ao posicionarmos a primeira janela no sinal de voz como representado pela janela J1 na Figura 5-a. Com isso, preenchemos o *buffer* ilustrado na Figura 5-b com as amostras da janela, calculamos e guardamos sua energia total E e apontamos um índice para a posição da primeira amostra, no caso $x(n)$. Assim, então, deslocamos a janela e subtraimos da energia total E a energia da amostra $x(n)$, adicionamos a energia da última amostra $x(n+N)$ da janela J2 à energia total E, guardamos seu valor na posição atual do *buffer* circular e deslocamos o índice para a próxima amostra $x(n+1)$. Repete-se até a janela percorrer todo o sinal de voz.

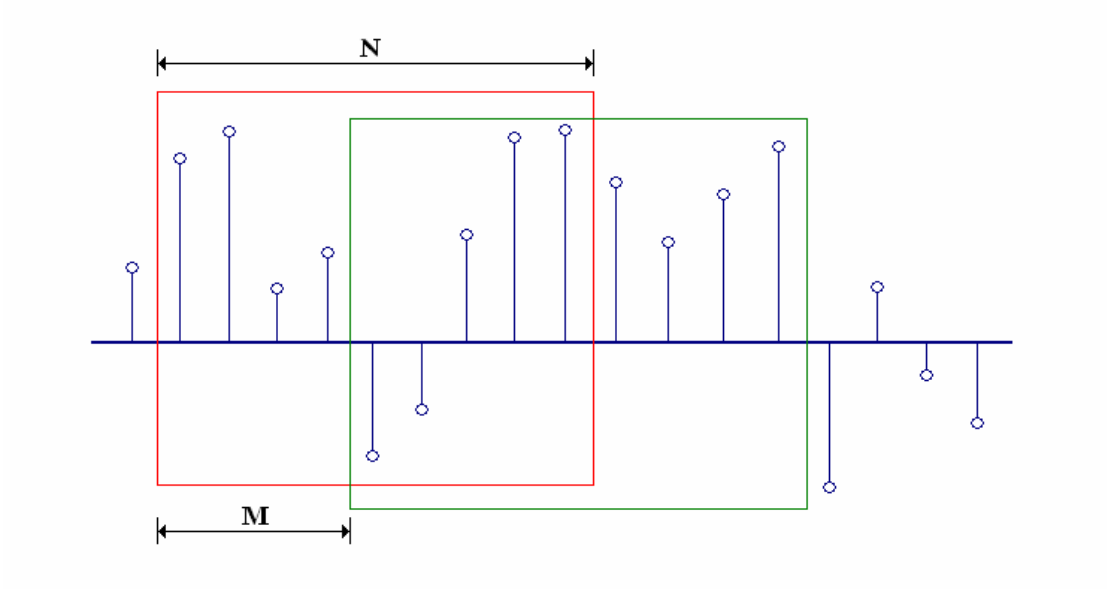

Figura 4 - Deslocamento de janela no domínio do tempo.

Note que sem utilizar o algoritmo *buffer* circular, a cada deslocamento de janelas tínhamos de realizar o cálculo de energia das N amostras da janela, enquanto que se o utilizarmos, precisamos apenas calcular as energias das amostras que estão entrando e saindo da janela, além das operações de adição e subtração.

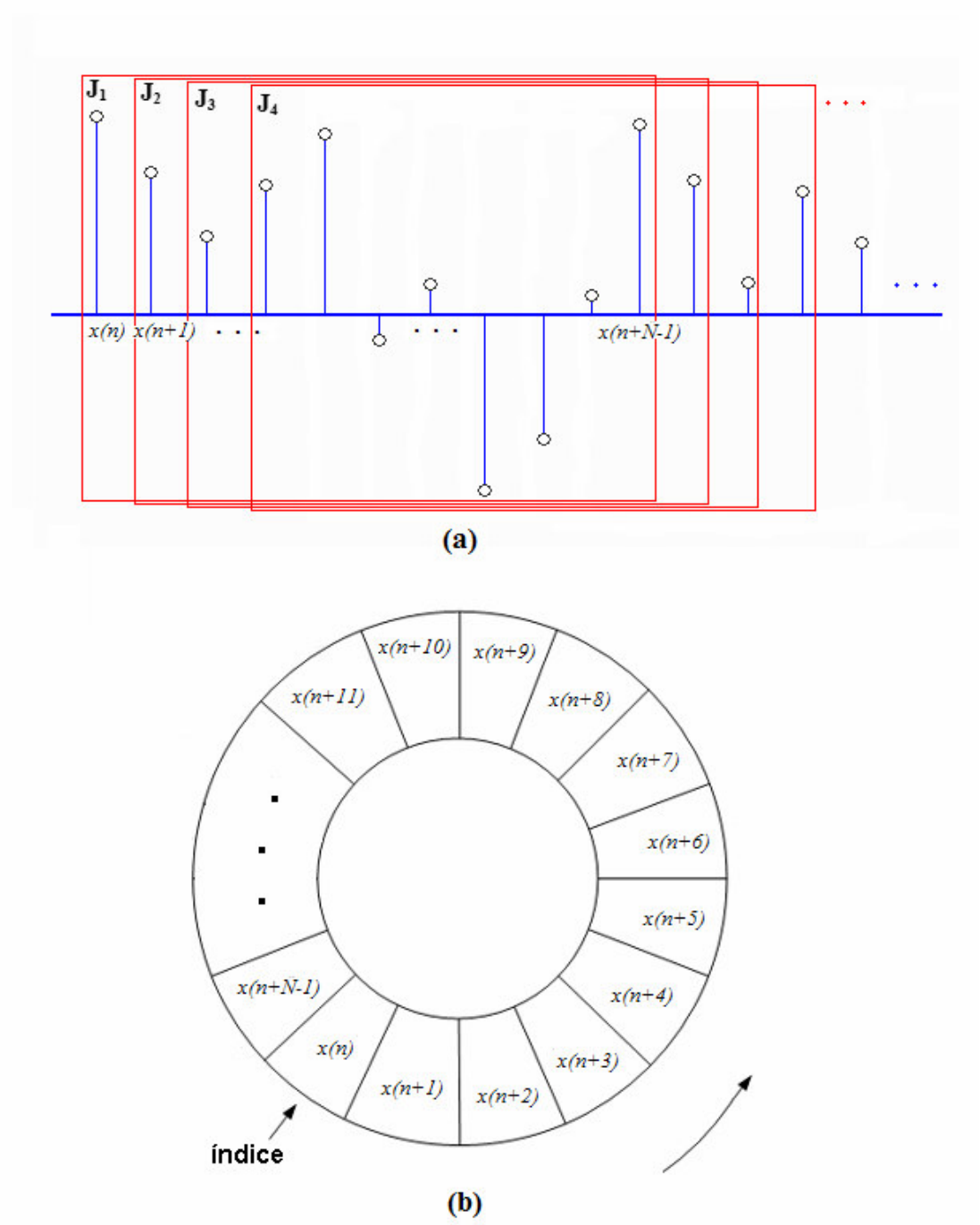

Figura 5 - Janela de tamanho *N* se deslocando de amostra em amostra em um sinal (a) e ilustração de um *buffer* circular.

2.1.3 Escalamento

Por dois motivos utilizaremos o escalamento para a geração dos sinais de informação. Primeiro para otimizar o custo de processamento e de memória, já que a energia de uma amostra com profundidade de 16 bits, por exemplo, terá 32 bits. Segundo para efeito de simplicidade e praticidade do sistema, que gerará sinais de

informação com as mesmas características do sinal original, tanto na taxa de amostragem quanto na quantidade de bits utilizados para representar a amostra.

Sabemos que a profundidade de bits para representar a amostra pode não ser suficiente para representar a energia. Por exemplo, calculando a energia amostra a amostra, como na equação (2.1), do trecho de sinal de áudio [5 6 9 14 16 10 3 -2 -3 -1] com 6 bits de profundidade, obtemos o trecho [25 36 81 196 100 9 4 9 1]. É fácil notar que as amostras do trecho de energia ultrapassam o máximo valor que uma amostra com profundidade de 6 bits, 0 a 63, pode alcançar.

$$E_S[E] = \frac{E}{S} \tag{2.3}$$

O objetivo do escalamento aqui consiste em dividir a energia E calculada por uma escala S onde o resultado $E_S[E]$ possa ser representado com a mesma quantidade de bits que a amostra $x(n)$. Então, considerando que a energia E tenha sido calculada a partir de uma amostra com seu valor máximo, a escala S deva ser também esse mesmo valor, pois, assim, o valor da energia escalada $E_S[E]$ seria também o valor máximo possível, ou seja, 2^{b-1}, como na equação (2.4), onde b é o número de bits que representam a amostra $x(n)$. Isso garantiria que todos os valores possíveis de amostras tenham uma energia que possa ser representada na mesma quantidade de bits das amostras originais.

$$S_{amostra} = \frac{2^{b-1}}{f_E} \tag{2.4}$$

Entretanto, sabe-se que sinais de voz raramente atingem o valor máximo possível, se concentrando abaixo da metade. Assim, na equação (2.4), é apresentado o fator de escalamento f_E para aumentar a faixa onde o sinal de informação gerado irá atingir. Obviamente que o fator de escalamento f_E deve ser definido precisamente para a amostra final não ultrapassar o valor máximo.

$$E_S[E[n]] = \frac{E[n] * f_E}{2^{b-1}} \tag{2.5}$$

A quantidade de bits que representam uma amostra, no caso do material à disposição, é b = 16 bits. A energia escalada $E_S[E(n)]$ deve ser o máximo valor aceitável para a amostra de energia, que no caso deve ser igual ao valor máximo do tipo de variável, 2^{b-1}. Após algumas análises de sinais de voz, encontramos o máximo do sinal igual a 11247. Porém, visando garantir uma margem razoável de segurança, determinamos $E(n)$ = 20000. Se aplicarmos os valores de b, $E_S[E(n)]$ e $E(n)$ na equação (2.5), encontramos o fator de escalamento f_E igual a $5{,}36 \times 10^4$.

Considerando o escalamento para o cálculo da energia de um segmento completo, precisamos ainda escalar o resultado pelo comprimento N do segmento, da forma indicada na equação (2.6).

$$E_S[E] = \frac{E * f_E}{2^{b-1} * N} \tag{2.6}$$

É claro que a determinação de bom momento ou não independe do escalamento, porém, é fácil perceber que sem o escalamento o custo de processamento e de memória é muito maior do que com escalamento.

2.2 Análise dos Métodos

Na Figura 6 vemos o exemplo do sinal de informação de um trecho de bom momento passado pelos métodos com janelamento superposto e não superposto com o tamanho da janela N igual a 250 ms, 500 ms e 1000 ms. É possível ver que nos dois casos, de superposição ou não, as janelas menores apresentam variações mais rápidas.

Apesar de em todos os casos ser fácil perceber que a energia da voz aumentou, não há como visualmente e em um único exemplo dizer qual o método e parâmetro irá ser mais eficiente na detecção de bons momentos. Assim, faremos análises estatísticas baseadas em sinais de informações gerados pelos métodos no Sinal I da base de dados apresentada anteriormente.

Para as análises estatísticas foram usados 1 sinal de informação gerado sem janelamento, 3 com janelamento superposto de deslocamento M = 1 e tamanho da

janela N = 250 ms, 500 ms e 1000 ms, e, por fim, mais 3 com janelamento não superposto de deslocamento $M = N$ = 250ms, 500ms e 1000ms.

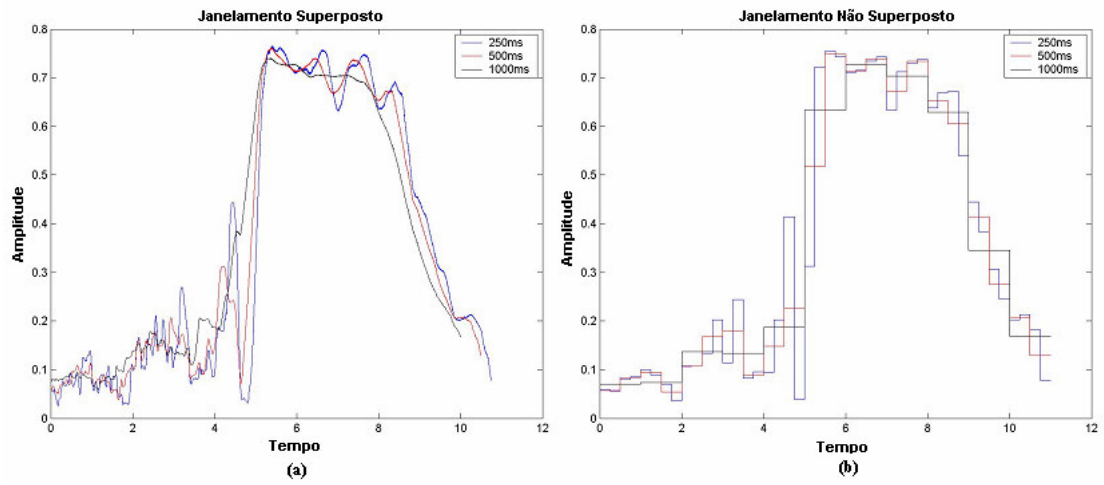

Figura 6 - Exemplo sinais de informação gerados com os métodos de janelamento superposto (a) e não superposto (b).

2.2.1 Primeiro Nível de Análise

Em um primeiro nível, faremos somente uma análise superficial baseada nos histogramas dos sinais de informação, separados em pontos marcados como bons momentos e pontos marcados como lances normais.

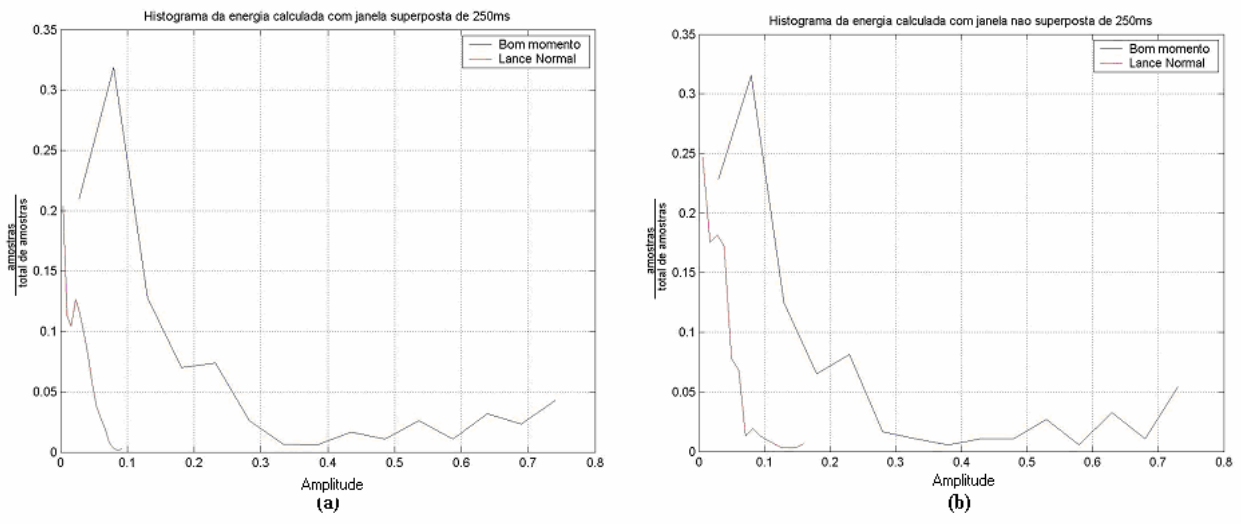

Figura 7 - Distribuição estatística do sinal de energia calculada com janela superposta (a) e não superposta (b) de 250ms.

Para todos os métodos com janelamento é fácil notar que a distribuição de bom momento se destaca da de lance normal, de M = 1 e $M = N$. Entretanto, também

é fácil reparar que a grande maioria dos pontos de bons momentos se encontra muito próxima à distribuição de lances normais. É possível ainda notar no histograma da Figura 9-a, com o método com janelamento superposto, que a distribuição de bons momentos se separa mais da distribuição de lances normais do que nos demais casos. A exceção é o caso sem janelamento, onde as distribuições são praticamente iguais e levemente deslocadas, e que se torna muito difícil determinar um limiar que possa classificar o trecho em questão.

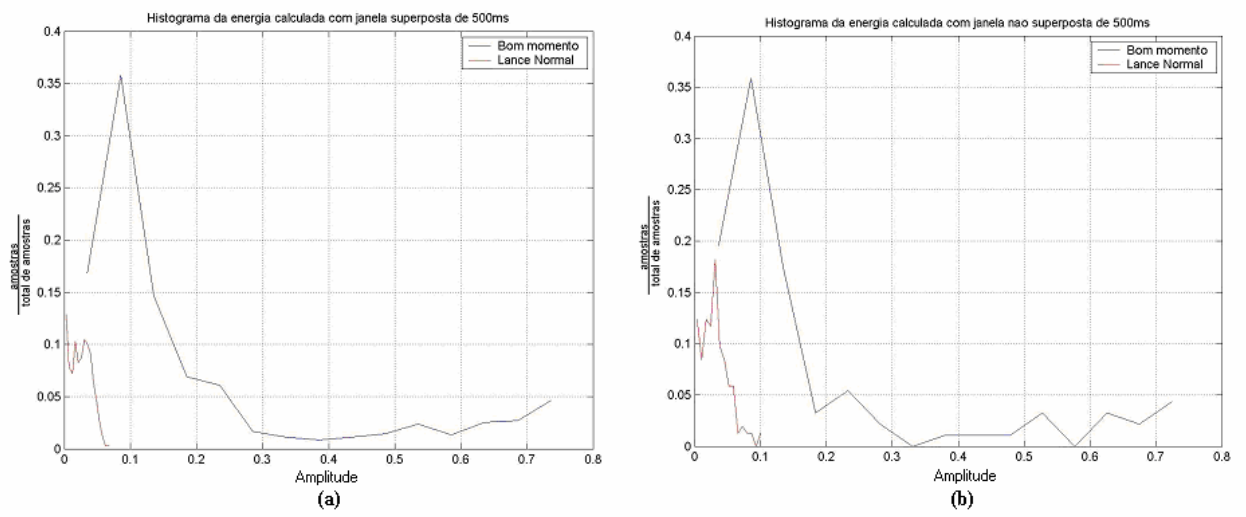

Figura 8 - Distribuição estatística do sinal de energia calculada com janela superposta (a) e não superposta (b) de 500ms.

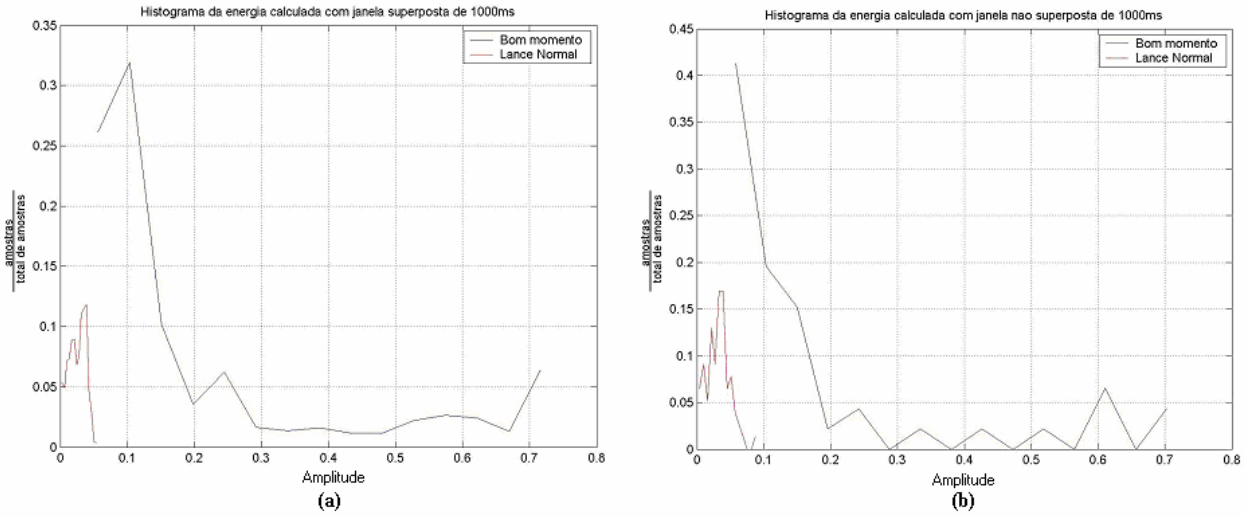

Figura 9 - Distribuição estatística do sinal de energia calculada com janela superposta (a) e não superposta (b) de 1000ms.

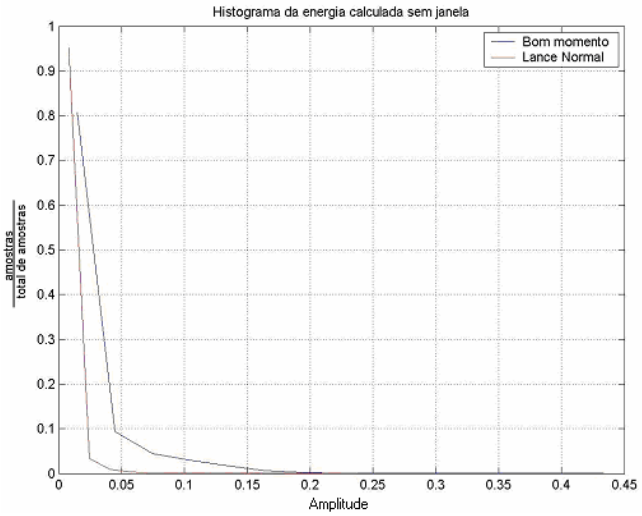

Figura 10 - Distribuição estatística do sinal de energia calculada sem janela.

Em um primeiro nível de análise, não é possível destacar nenhum método e/ou parâmetro como mais eficaz para a determinação de um limiar, tornando necessário um segundo nível de análise. Porém, ficou claro que o método do sinal de informação sem utilizar nenhum janelamento pode ser descartado para as análises subseqüentes.

2.2.2 Segundo Nível de Análise

Visto que o primeiro nível foi baseado na análise visual dos histogramas e poucas conclusões puderam ser extraídas, o segundo nível procurará quantificar o comportamento de cada variável.

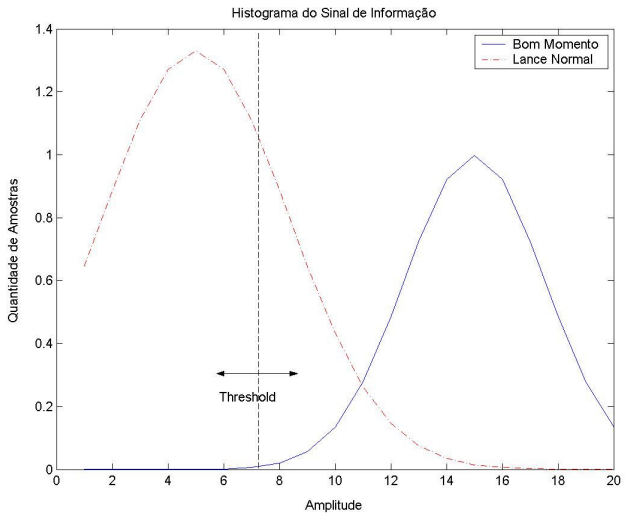

Figura 11 - Exemplo de análise do histograma do sinal de informação.

A partir dos histogramas anteriores criados no primeiro nível de análise, podemos variar o limiar de classificação ('*threshold*') e traçar qual o percentual dessa mesma distribuição está sendo definida corretamente, já que os pontos de lance normal devem ficar à esquerda do limiar enquanto os de bom momento à direita, como exemplifica a Figura 11. Por exemplo, quando o limiar for zero, 100% dos pontos de bons momentos estarão sendo marcados corretamente e 0% dos pontos de lances normais estará correto. No outro extremo, se o limiar for o máximo valor encontrado, 0% dos pontos de bons momentos estará correto enquanto 100% dos pontos de lances normais estarão corretos.

O interessante dos gráficos resultantes, nas Figuras 12 e 13, é notar em que ponto há o máximo percentual de acerto tanto para bons momentos quanto para lances normais.

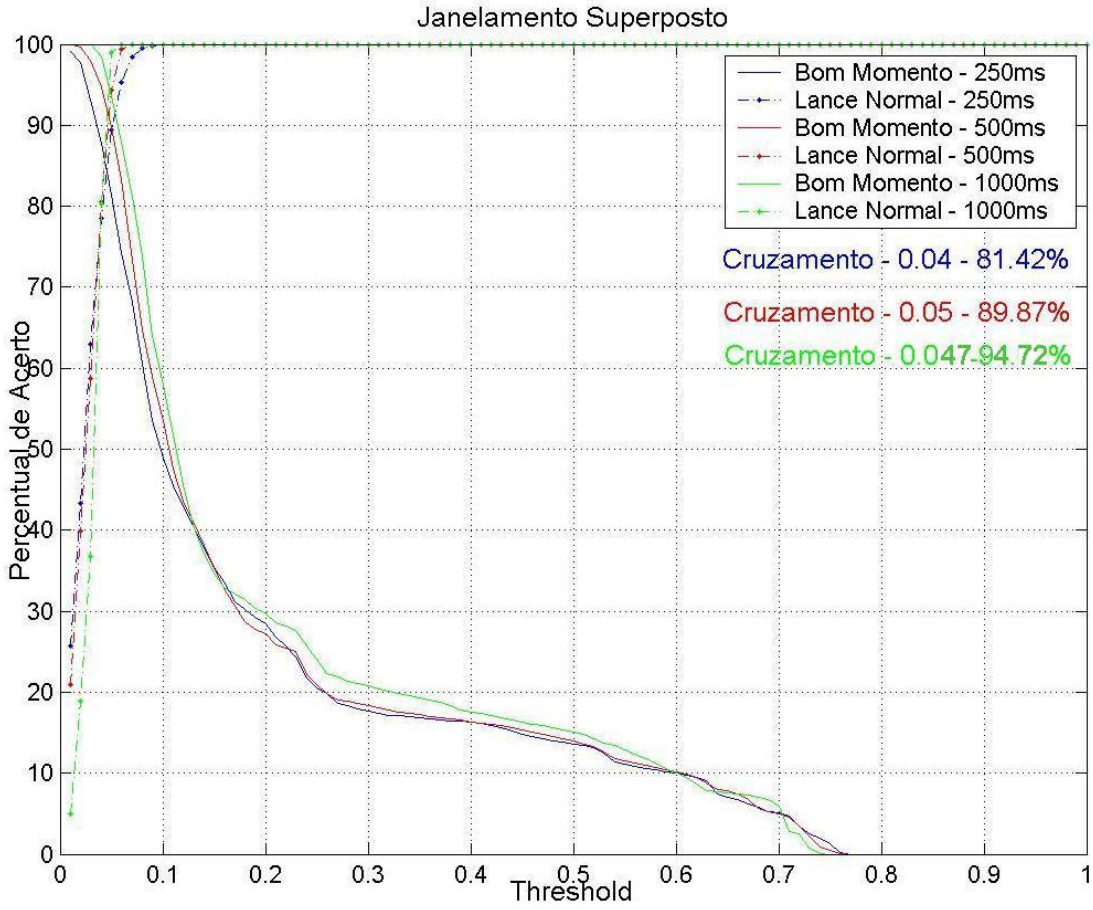

Figura 12 - Curvas de percentual de acerto pelo limiar para o janelamento superposto com os três tamanhos de janela *N*.

No caso do janelamento superposto, demonstrado na Figura 12, a janela de tamanho N = 1000 ms apresenta o cruzamento com o maior percentual de acerto, 94.72%. A Figura 13 ilustra o gráfico de percentual de acerto pelo limiar para o método com janelamento não superposto, onde mais uma vez a janela N = 1000 ms apresentou o melhor resultado. O cruzamento entre lances normais e bons momentos ocorreu em 91.30% de acerto na classificação.

Nos dois métodos, janelamentos superposto e não superposto, podemos ver que a janela com tamanho N = 1000ms apresentou um melhor percentual de acerto, sendo que o janelamento superposto superou um pouco o janelamento não superposto.

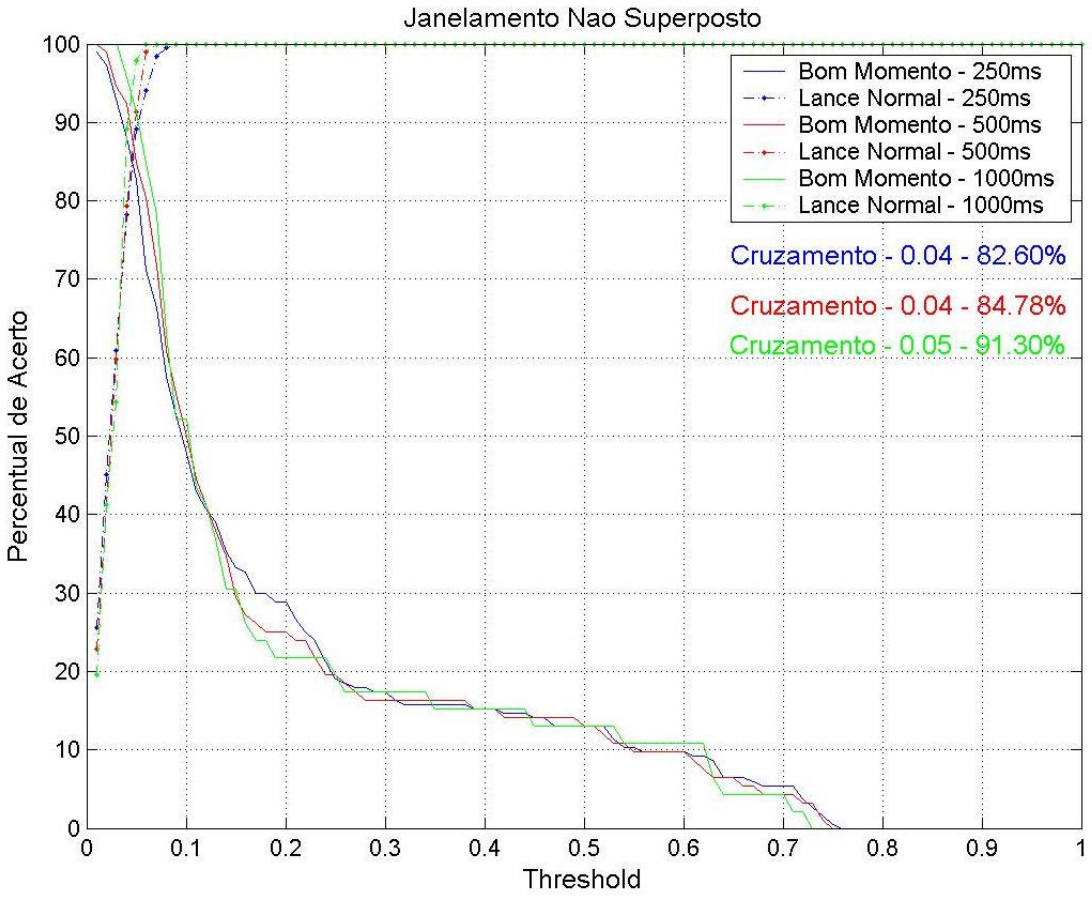

Figura 13 - Curvas de percentual de acerto pelo limiar para o janelamento não superposto com os três tamanhos de janela *N*.

Porém, se o limiar for definido com o valor do cruzamento ainda haverá pontos de bom momento que serão marcados como lance normal. Assim, é interessante que o limiar seja deslocado para a esquerda para que o percentual de

acerto dos bons momentos aumente, mesmo que o dos lances normais diminua, pois é mais crítico que o sistema deixe de marcar um bom momento do que marque um lance normal a mais. Portanto, baseado no argumento anterior e na Figura 14, o limiar foi definido em 0.04, com o percentual de acerto para os bons momentos atingindo 98.40% e dos lances normais 80.35%. Além disso, o segundo nível de análise determinou também que o método de geração do sinal de informação deva ser com janelamento superposto de deslocamento M = 1 amostra e janela de tamanho N = 1000 ms.

Evidentemente que o limiar foi encontrado fazendo análises sobre o material de um jogo específico, o que não significa que ele valerá para todos os outros. Assim, tanto o método, quanto seus parâmetros deslocamento M, tamanho N e limiar de classificação podem ser aperfeiçoados mais a frente, em capítulos que tratarão da validação do método empregado.

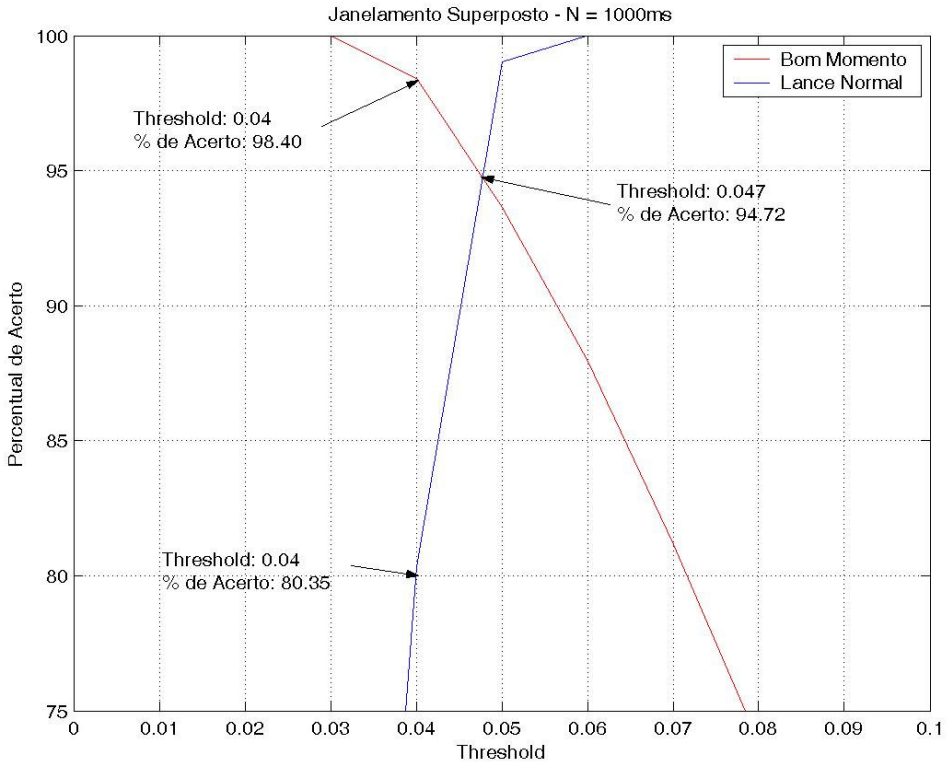

Figura 14 - Detalhe do cruzamento entre a curva de Bom Momento e Lance Normal.

Sabemos que todo o desenvolvimento do método de geração do sinal de informação foi feito utilizando somente um locutor como base, e, então,

provavelmente o limiar encontrado seja dependente do locutor. Assim, seria interessante que se o método fosse aplicado a outro locutor houvesse o ajuste deste limiar. Por exemplo, se o narrador tivesse a voz com volume maior o limiar aumentaria, caso contrário, diminuiria.

2.3 Conclusões

Nesse capítulo, foi apresentado o embasamento teórico necessário para entender como foram desenvolvidos os métodos para gerar sinais informação baseados na energia que fossem interessantes para um sistema de detecção de bons momentos de programas esportivos.

Alguns parâmetros foram avaliados como N (comprimento da janela) e M (deslocamento da janela). Desenvolvendo um método com parâmetros variados, testando-os, e determinando um limiar bem caracterizado, é possível afirmar com maior certeza que há grandes chances do sistema ter sucesso na detecção de bons momentos.

Após a apresentação, os métodos foram discutidos, assim como suas implementações e parâmetros necessários. Assim, foi possível implementar as rotinas para gerar os sinais de informação que serviram de material para a análise e posterior definição do método, dos parâmetros e do limiar de classificação, que pode ser ajustado caso o narrador seja trocado.

O próximo capítulo discutirá outra maneira de gerar o sinal de informação baseado no *pitch* da voz.

Capítulo 3 - Sinal de Informação Baseado no *Pitch*

Além de determinar os trechos de bons momentos baseados em sinais de informação de energia, seria muito interessante realizar outro caminho para enriquecer o sistema e torná-lo mais robusto.

Este capítulo realizará estudos e análises sobre o *pitch* para o desenvolvimento de um método capaz de gerar sinais de informação de bons momentos.

Na seção 3.1 será estudado o comportamento do *pitch* nos sinais de narração real e um método capaz de gerar um sinal contendo informações que caracterizem os bons momentos. A seção 3.2 tentará, a partir dos resultados obtidos com o estudo da

seção 3.1, apontar os parâmetros mais indicados para a classificação de bom momento ou não. Por fim, a seção 3.3 resumirá o capítulo e discutirá os resultados obtidos.

3.1 Formulação do Problema

O período de *pitch Tp* é provocado pelos movimentos quase periódicos das cordas vocais na faringe e é o inverso da freqüência fundamental percebida pelo sistema auditivo humano, como afirma [1]. A Figura 15 ilustra um sinal ideal dos batimentos das cordas vocais.

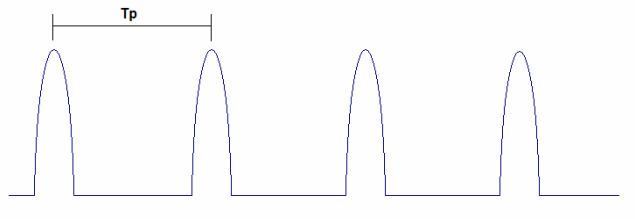

Figura 15 - Período de *pitch* em um suposto sinal de batimento.

O algoritmo para extrair a freqüência fundamental de um sinal de voz explora a periodicidade do sinal de voz. Assim, mesmo com um sinal com diversas componentes frequenciais, [6] diz que se realizarmos a autocorrelação R_{xx}, demonstrada em [3], de um sinal de voz $x(n)$ com *lag* $\tau = 0$, chegaremos ao sinal de autocorrelação de onde podemos determinar a periodicidade do sinal, onde

$$R_{xx}(\tau) = \sum_{n} x_n \overline{x_{n-\tau}} \,. \tag{3.1}$$

[5] e [6] dizem que calculando o intervalo entre os maiores picos da autocorrelação $R_{xx}(\tau)$ teremos uma estimativa do período de *pitch* T_p, como mostra a Figura 16.

Porém, só devemos calcular o *pitch*, em trechos sonoros, pois, segundo [6], em trechos surdos não há freqüência fundamental da voz, já que as cordais vocais não se movimentam. Com isso vemos na Figura 17 que não há como encontrar dois picos na autocorrelação para calcular o *pitch*, o que pode se tornar fonte de erros no sinal de informação.

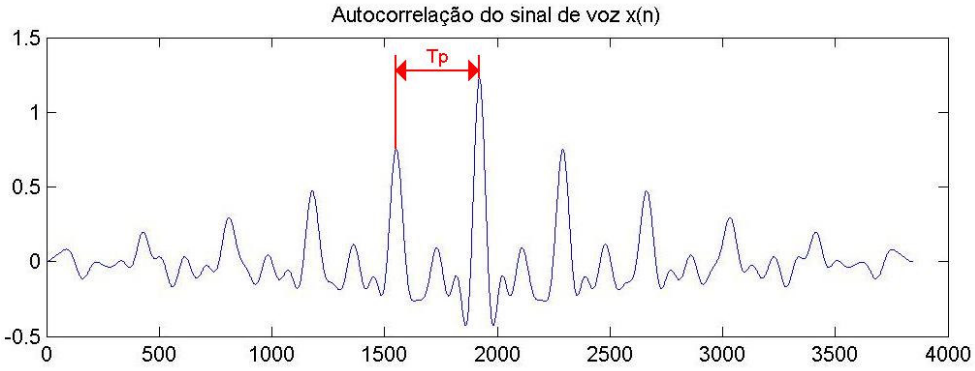

Figura 16 - Ilustração de como obter o período de *pitch* de um sinal de autocorrelação R_{xx}.

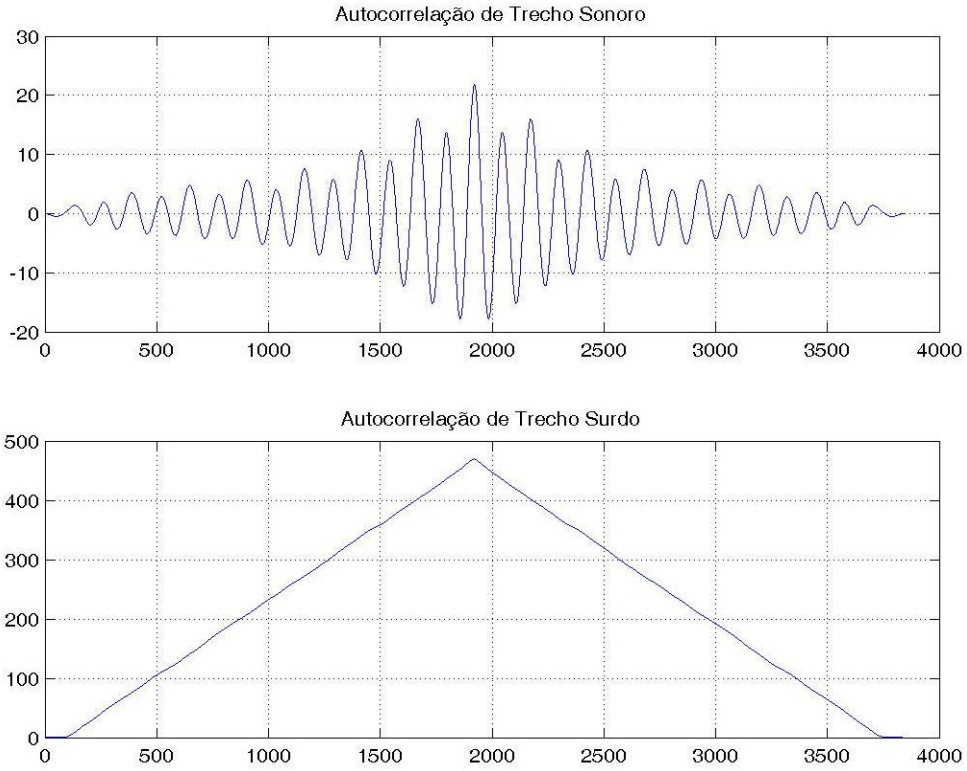

Figura 17 – Autocorrelação de trecho sonoro e surdo.

Desta maneira, estudaremos mais tarde se os trechos surdos influenciam suficientemente para que sejam excluídos do cálculo quando ocorrerem. Em contrapartida, a autocorrelação dos trechos sonoros possibilita a extração do *pitch*. A Figura 18 ilustra duas situações do uso da autocorrelação em sinais sonoros, onde é possível notar que os picos de maior amplitude estão mais próximos na situação de gol do que na situação de lance normal, ou seja, a voz no lance de gol teve uma freqüência fundamental maior do que no lance normal. Assim, exploraremos este fato para classificar um lance qualquer.

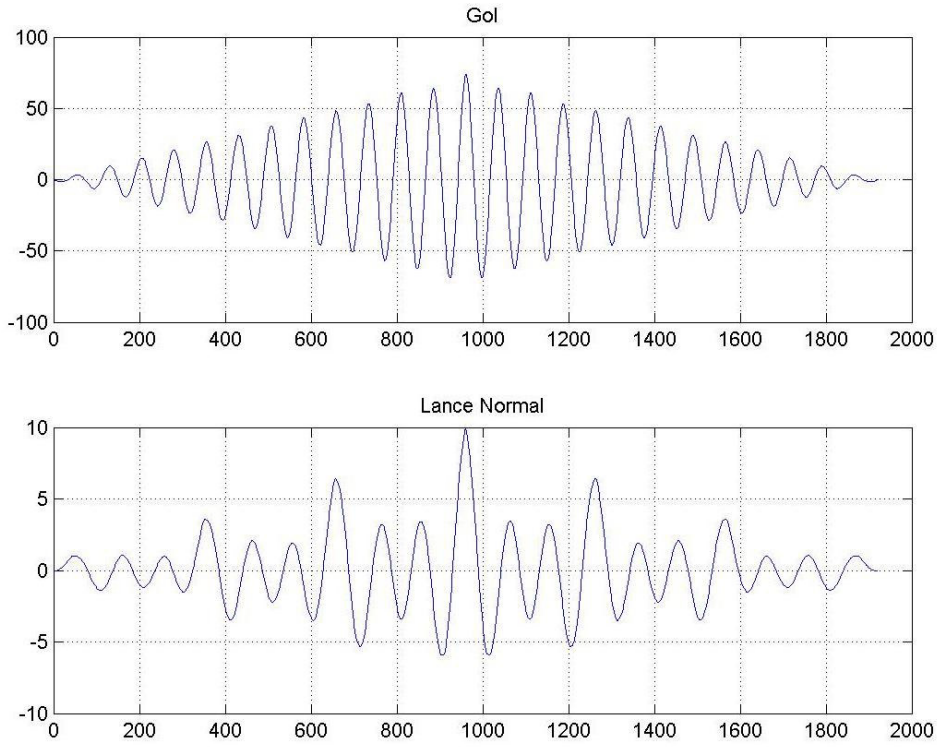

Figura 18 - Autocorrelação para um trecho com Gol e outro com Lance Normal.

Sabendo que a correlação se trata de uma convolução, o custo de processamento dessa operação é muito alto quando utilizamos janelas do tamanho que será proposto mais a frente, tornando o processo lento. Portanto, com intuito de torná-lo mais veloz, [4] diz que pode-se utilizar o domíno da freqüência, de modo que

$$R_{xx}(\tau) = IDFT\{|\, DFT[x(n)]\,|^2\}, \tag{3.2}$$

onde *DFT(.)* e *IDFT(.)* indicam transformada de Fourier direta e inversa, respectivamente.

A freqüência fundamental da voz masculina está em torno de 150 Hz, em geral, acima dos 100 Hz. Assim um período de *pitch* T_p será no máximo de 10 milissegundos. A fim de realizar o cálculo do *pitch* pela autocorrelação de forma mais precisa, seria interessante ter mais de três ciclos no sinal de voz. Sabendo disso, poderíamos, então, utilizar janelas maiores que 30 milissegundos que as condições seriam satisfeitas.Portanto, para garantir uma pequena margem de segurança utilizaremos janelas de 40 milissegundos. Além disso, pensando em evitar

interferências provenientes de outras fontes, será aplicado antes de qualquer cálculo um filtro passa-baixas limitando a banda em 1 kHz.

Para evitar cálculos desnecessários, antes de calcular qualquer trecho, o algoritmo identifica se o trecho em questão possui voz ou é essencialmente de silêncio. Isso é realizado com um simples cálculo da energia total do segmento, onde se um limiar for ultrapassado, trata-se como voz, se não, como silêncio.

Na geração do sinal de informação baseado no *pitch*, a cada segmento com voz é associado o valor da freqüência fundamental encontrada, e a cada trecho de silêncio é associado o valor zero. Em um primeiro momento trataremos os trechos surdos como sonoros, ou seja, tem o *pitch* calculado, apesar de ser esperado um erro. Com as análises subseqüentes, veremos se esses erros são prejudiciais ao método, ou se é suficientemente irrelevante para convivermos com eles.

3.2 Análise do Método

Com a apresentação dos fundamentos teóricos necessários ao entendimento, podemos realizar análises sobre o método para constatar se ele é capaz ou não de determinar em que trechos se encontram bons momentos.

Antes de analisar os resultados do método, é necessário encontrar o limiar que determina se o segmento que terá o *pitch* calculado é de voz ou silêncio, para que o método saiba se pode o descartar. Para isso, basta realizar uma simples análise baseada em histogramas de trechos marcados manualmente como voz e silêncio, para, a partir disso, encontrar um limiar satisfatório. Na Figura 19, nota-se a separação entre os histogramas, o que torna possível dizer que um trecho com energia menor do que 0.1 claramente trata-se de silêncio.

Com o método e seus parâmetros definidos, podemos usá-los e ver seus resultados. A Figura 20 ilustra um sinal de informação baseado no *pitch* gerado com o método proposto. Nele conseguimos ver claramente o gradual aumento da freqüência fundamental acompanhando a chegada do bom momento no sinal original, onde, pelo menos visualmente, já é possível ver que o método gera um sinal

com as informações desejadas. Porém, é impossível determinar, sem uma análise apurada, o limiar que consiga separar os bons momentos dos demais.

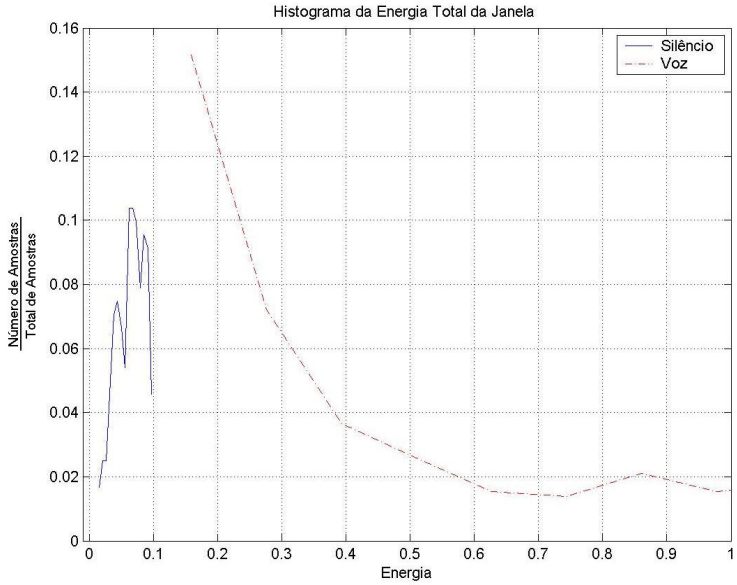

Figura 19 - Histogramas de energia para trechos de silêncio e de voz.

Então, assim como nas análises do capítulo anterior, faremos um primeiro nível de análise baseada em histogramas provenientes de trechos previamente marcados como bons momentos e lances normais no Sinal I. Estes dados são os mesmos usados no capítulo anterior.

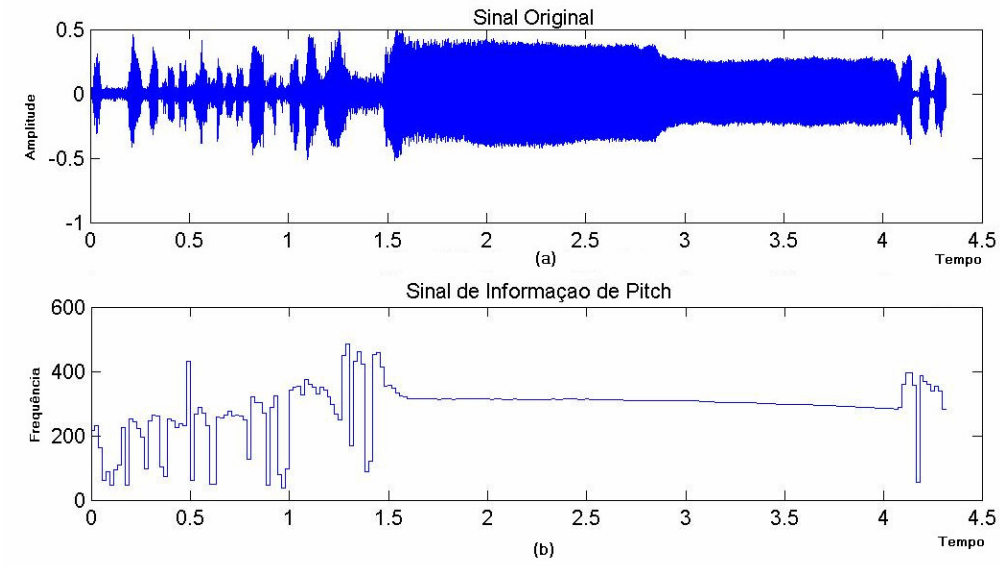

Figura 20 - Sinal Original de Voz (a) e o Sinal de Informação de *Pitch* (b) gerado a partir dele.

Com os histogramas ilustrados na Figura 21, é fácil observar que a distribuição dos trechos de bom momento fica bem separada da distribuição de trechos de lance normal, confirmando o que havíamos previsto na Figura 20. Além disso, já é bem mais fácil ver em qual valor da freqüência fundamental da voz que passamos a encontrar mais amostras de bons momentos.

As distribuições se cruzam em cerca de 225 Hz, porém é fácil notar que abaixo desse valor ainda existem muitas amostras de bom momento, como também acima dele há valores de lance normal. Porém, sendo preferível marcar alguns lances normais como bom momento do que deixar de marcar bons momentos, é interessante a redução desse limiar para 200 Hz. Desse modo, o sistema marcará essencialmente 87,5% dos pontos de bons momentos e 3,5% dos pontos de lances normais, porém isso pode ser facilmente resolvido criando um sistema em que seja possível que o operador descarte lances marcados que ele julgue errados.

Nota-se que mesmo sem realizar o descarte de trechos surdos, conseguimos chegar a um limiar que separe bem os bons momentos dos lances normais. Portanto, não há a necessidade imediata de aumentar a complexidade do método inserindo o cálculo de sonoridade dos trechos.

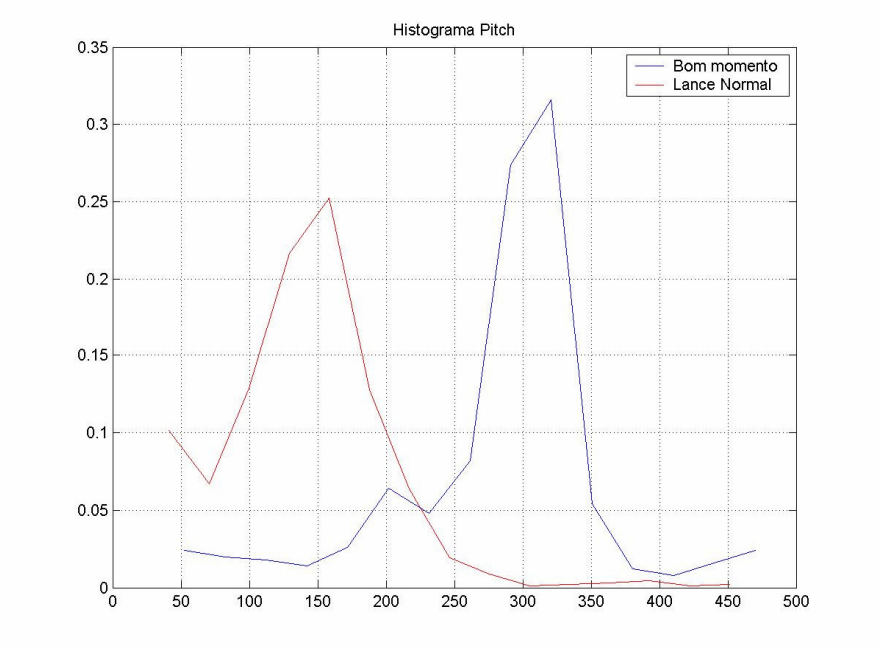

Figura 21 - Quantidade de amostras pela freqüência fundamental do sinal de informação baseado no *pitch*.

Assim como na geração do sinal de informação baseado na energia, podemos também ajustar o limiar caso o locutor seja diferente do Narrador I. Por exemplo, se o narrador tivesse a voz mais aguda o limiar aumentaria, caso contrário, diminuiria.

3.3 Conclusões

Neste capítulo vimos o desenvolvimento de mais um método para gerar o sinal de informação útil à detecção de bons momentos em um programa esportivo, o que traz ao sistema de detecção agora a possibilidade de seguir dois caminhos, em série ou em paralelo, para tentar obter informação de bom momento ou não, um pela variação da energia e outro pela variação da freqüência fundamental da voz.

Diferentemente do capítulo anterior, onde mais de um método e parâmetro foi discutido, aqui chegamos a definição do método muito mais rápido, pois havia menos parâmetros. Além disto, usamos as conclusões obtidas anteriormente.

Assim como na análise de sinais de informação baseados na energia, fizemos análises baseadas em histogramas das distribuições de trechos anteriormente marcados manualmente. Com a experiência adquirida ao longo do capítulo anterior, o primeiro nível de análise desses histogramas já foi suficiente para encontrar um limiar satisfatório para a freqüência fundamental, que pode ser ajustado caso o narrador seja alterado.

Após a definição de dois métodos capazes de gerar sinais de informação de onde possam ser extraídas informações de bons momentos, o próximo capítulo terá a missão de estudar um módulo de decisão que seja capaz de analisar os dois sinais gerados e decidir se deve ou não marcar o trecho como bom momento.

Capítulo 4 - Módulo de Decisão

Os sinais de informação dos capítulos 2 e 3 nos deram base para detectar trechos de bom momento em uma narração. Porém, por cada um ter seguido um caminho diferente, inevitavelmente, irão acontecer situações em que um sinal aponta para bom momento e o outro não, ou ambos apontam, mas em pontos diferentes Isto nos leva a necessidade de um módulo que seja responsável por decidir se será ou não um bom momento, a partir de quando, e até quando. Além disso, o módulo tem de ser capaz de identificar bons momentos muito próximos, percebendo que há apenas um trecho, e de descartar trechos muito curtos, que podem ter sido inicialmente classificados erroneamente como bom momento.

Dessa forma, a seção 4.1 estudará a marcação de bons momentos a partir de dois sinais de informação, sendo a determinação da região de bom momento, para posterior determinação de início e fim. A seção 4.2 tratará a união de bons momentos

vizinhos que sejam o mesmo. A seção 4.3 discutirá o descarte de trechos identificados como bons momentos, mas que sejam curtos. Por fim, a seção 4.4 resumirá o capítulo e comentará os resultados obtidos.

4.1 Decisão por Bom Momento

O primeiro passo do módulo de decisão é realizar uma marcação instantânea dos trechos a partir dos limiares determinados no desenvolvimento dos métodos de geração do sinal de informação. No Capítulo 2, vimos que $E = 0.04$ em uma janela de comprimento $N = 1000$ ms seria adequado para separar os bons momentos dos lances normais. Já no Capítulo 3, vimos que trechos que tivessem a média da freqüência fundamental maior que 200 Hz poderiam ser considerados como bons momentos independentemente do comprimento N da janela. Assim, para utilizarmos a mesma janela para marcar tanto o sinal baseado na energia quanto o baseado no *pitch*, utilizaremos uma janela de comprimento $N = 1000$ ms, mesmo sabendo que a janela utilizada para determinar o limiar de *pitch* era 25 vezes menor. Além disso, a janela será deslocada de $M = 500$ ms para aumentar a quantidade de marcas.

A Figura 22 ilustra um exemplo de marcação instantânea dos sinais de informação baseada simultaneamente na energia e no *pitch*. É fácil perceber que os dois limiares foram satisfatórios para que na região de bom momento houvesse trechos marcados, porém, podemos notar que nem todos os pontos do sinal da Figura 22-b estão na Figura 22-c e vice-versa.

Após determinar marcações instantâneas para ambos os sinais de informação, temos que avaliá-las para definir um sistema que possa definir onde há, se inicia e termina o bom momento. Se gerarmos as marcações Sinal I, teremos material suficiente para avaliar o perfil das marcações em cada sinal de informação.

A primeira característica observada nas marcações é que o sinal de informação baseado no *pitch* gera muito mais marcas do que o baseado na energia, e muitas delas em trechos onde não há bom momento, o que gera certa desconfiança na eficácia das marcações feitas pelo *pitch*. Enquanto isso, o sinal baseado na energia é mais eficiente ao marcar, pois a grande maioria das marcas está em bons momentos. Mas por outro lado, outra característica observada é que o sinal baseado na energia

demora mais a marcar um bom momento do que o baseado no *pitch*, o que poderia criar um problema na determinação do início. Dessa forma, poderíamos ter um sistema que buscasse pelas regiões de bom momento através das marcações do sinal de informação baseado na energia, para posteriormente confirmar e definir seus limites com as marcações do sinal de informação baseado no *pitch*.

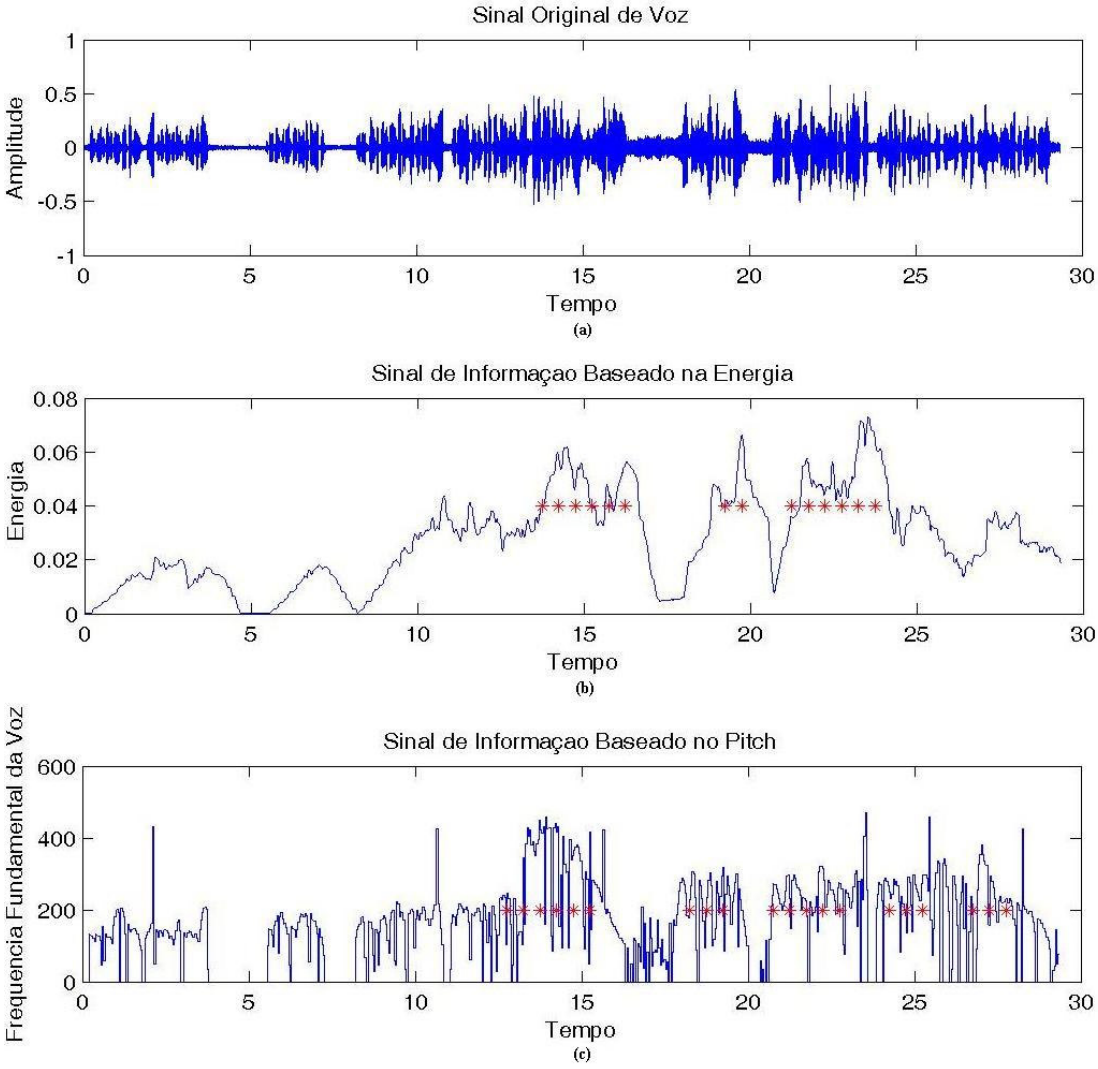

Figura 22 - Sinal de informação baseado na energia (b) e sinal de informação baseado no *pitch* (c) com bons momentos marcados. Ambos gerados a partir do sinal de voz (a).

4.1.1 Busca pela Região de Bom Momento

Mesmo sabendo que há bem menos marcações de bom momento no sinal de informação baseado na energia do que no baseado no *pitch*, ainda assim existem marcações mal feitas. Portanto, para evitar perda de tempo com marcações irrelevantes, temos que tentar descobrir quantas marcações em seqüência são

suficientes para caracterizar um bom momento. Desse modo, ainda acontecerão marcações indesejadas, porém bem menos freqüentes.

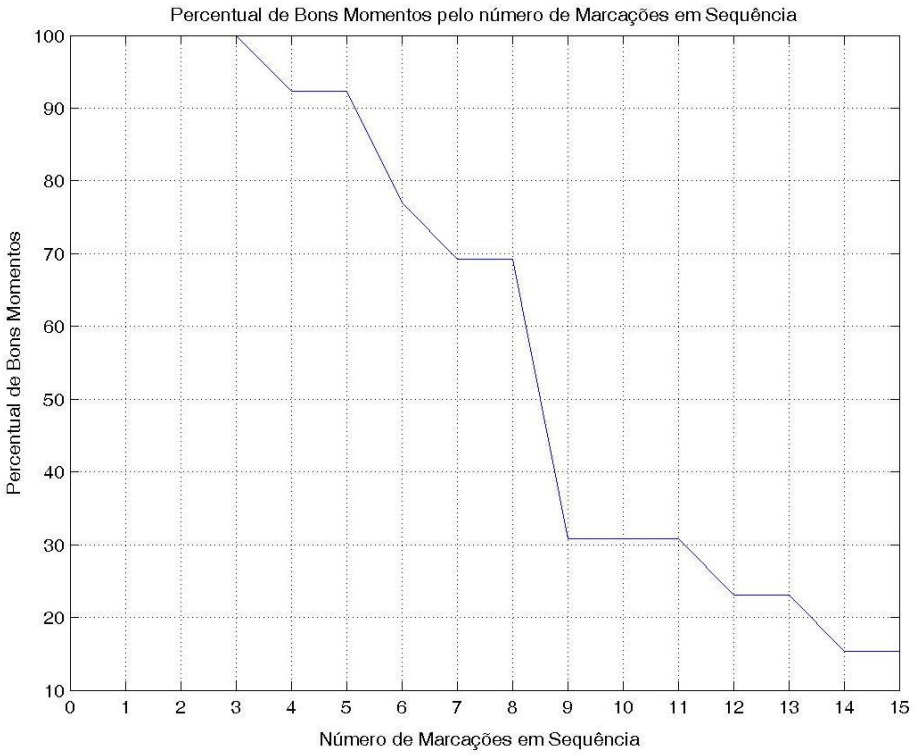

Figura 23 - Percentual de Bons Momentos atendidos pelo número de amostras em seqüência.

No gráfico da Figura 23, vemos que à medida que o número de marcações em seqüência vai aumentando o percentual de bons momentos atendidos vai diminuindo. Porém, sabemos também que quanto menor o número de marcações em seqüência, mais pontos indesejados serão considerados. Sendo assim, vislumbrando diminuir a marcação de pontos indesejados, mas ao mesmo tempo atender a todos os bons momentos, utilizaremos o maior número de marcações em seqüência possível, que é três.

4.1.2 Determinação de Início e Fim

Foi observado que em nenhuma das vezes que houve uma seqüência de três ou mais marcações do sinal de informação baseado na energia em trechos indesejados, havia marcações do sinal de informação baseado no *pitch*. Isto nos leva a outro passo para reduzir a marcação de pontos indesejados, que é verificar a presença de marcas do sinal baseado no *pitch* na mesma região. Para bons momentos,

as marcações de *pitch* sempre aparecem espalhadas pela região, fato que será explorado para determinação de início e fim.

Na Figura 22 vemos um comportamento interessante para a detecção do início do trecho desejado. Antes do instante em que o sinal de informação baseado na energia aponta o início de um bom momento, o sinal baseado no *pitch* já marcou pontos, isto aconteceu em 88% das vezes no Sinal I. Isto faz sentido, já que vimos no Capítulo 2 que o sinal de informação baseado na energia varia de forma muito mais lenta do que o baseado no *pitch*. Dessa maneira, ao sabermos que existe um bom momento pela energia e verificando o aumento da freqüência fundamental da voz momentos antes, podemos marcar o início do bom momento.

Ainda na Figura 22 é possível ver que ao fim da primeira seqüência de marcações do sinal baseado na energia não há marcações no sinal de *pitch*, enquanto que ao fim da última seqüência há. Ao contrário da detecção do ponto de início, o sinal baseado no *pitch* pouco ajuda para detectar o fim do bom momento, pois em apenas 36% das vezes houve marcas de *pitch* após as marcas da energia. Entretanto, mesmo com índice baixo de ocorrência, quando existir é interessante usá-lo para determinar o fim, quando não, o fim permanece ao término das marcas do sinal de informação baseado na energia.

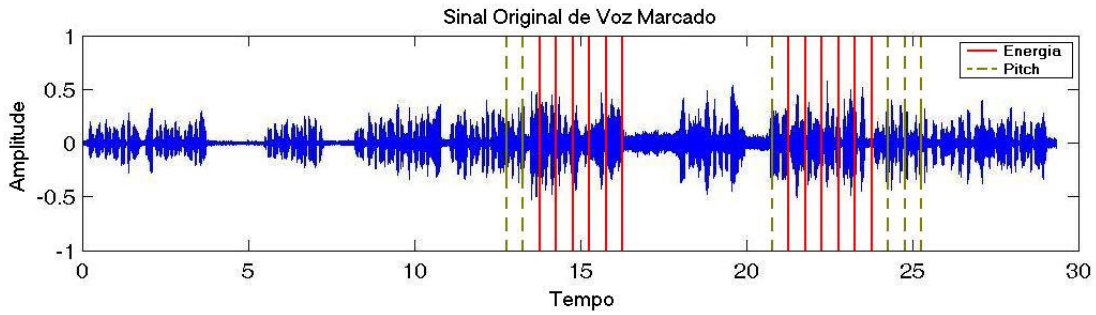

Figura 24 - Sinal original de voz com as marcas de energia e de *pitch* que serão consideradas pelo sistema.

Se aplicarmos os passos descritos aos sinais da 22, chegaremos à marcação dos bons momentos como indicado na Figura 24. Note que o trecho marcado pela energia com apenas duas marcações em seqüência foi descartado. Também podemos perceber que em ambos os trechos marcados pela energia, o *pitch* foi útil para

determinar o início do bom momento. Porém, apenas no último trecho ele foi utilizado para determinar o fim.

4.2 União de Trechos Marcados

O próximo passo após obter os trechos é uni-los quando fizerem parte do mesmo bom momento. Por exemplo, na Figura 24, temos dois trechos muito próximos que possivelmente pertencem ao mesmo bom momento, pois é fácil observar que menos de cinco segundos os separam. Dessa forma, o sistema indicaria dois trechos, enquanto na realidade trata-se de um só.

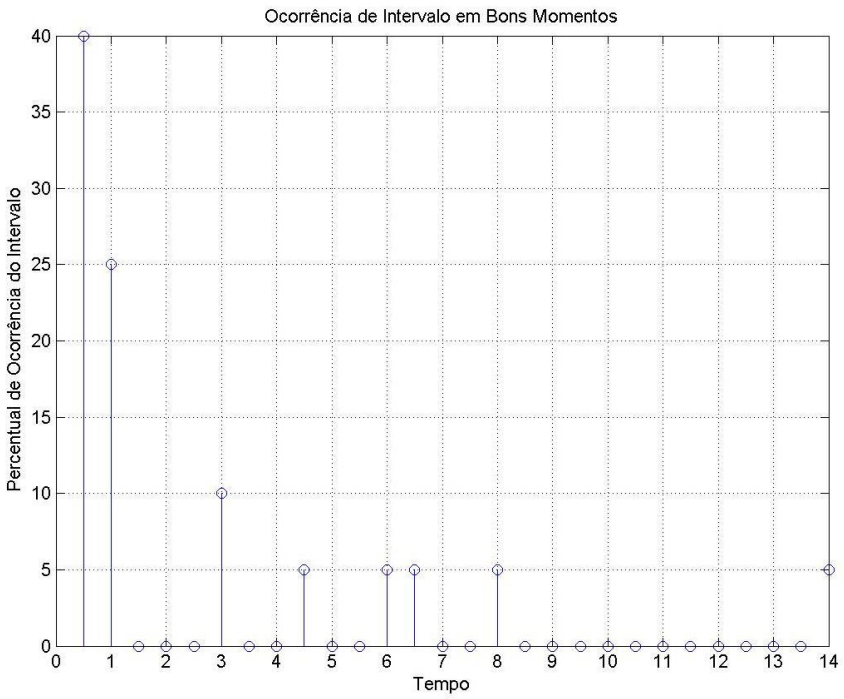

Figura 25 - Percentual de ocorrência do intervalo em segundos entre trechos de um mesmo bom momento.

Assim, se realizarmos um estudo nos trechos gerados através do Sinal I podemos analisar os intervalos de tempo entre trechos que pertencem ao mesmo bom momento. A Figura 25 exibe o gráfico que expressa essa análise, onde podemos ver que somente 5% dos intervalos foram maior que oito segundos e que mais de 80% têm intervalo menor que cinco segundos. Dessa maneira, se estipularmos um intervalo de dez segundos, atenderemos a praticamente todos os casos, fazendo com que o exemplo da Figura 24 apontasse somente um bom momento ao invés de dois.

Aplicando o método no Sinal I sem a união de trechos vizinhos, obtivemos 53 trechos marcados como bons momentos. Enquanto que aplicando com a união, temos 24 trechos marcados. Portanto, a união de trechos marcados que estejam a menos de dez segundos um do outro, trouxe uma melhoria significativa ao método.

4.3 Descarte de Trechos Curtos

Outra decisão que o módulo pode ser capaz de tomar é excluir trechos muito curtos que o método possa ter marcado. Porém antes de fazer isso, temos de estudar se esse fato realmente é uma fonte de erro, e se pode acontecer de um trecho muito curto ser um bom momento, o que seria altamente prejudicial ao método.

Sabemos que o menor trecho que o módulo considera como bom momento é o que tem três marcações em seqüência no sinal de informação baseado na energia, o que significa um segundo e meio de energia alta. Podemos, então, observar os trechos marcados e verificar quantos deles são bons momentos e pouco maioresa que um segundo e meio, tal como três segundos.

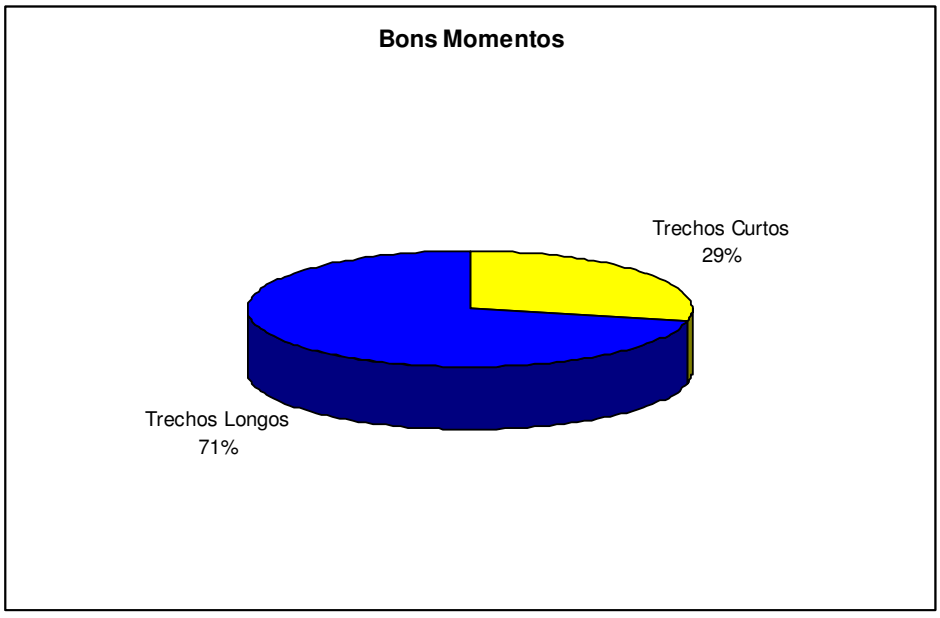

Figura 26 - Bons Momentos distribídos entre trechos curtos e longos.

No gráfico da Figura 26, é fácil notar que há bons momentos que são trechos curtos, o que inviabiliza o descarte. Mesmo sabendo que a maioria dos bons

momentos são longos, o método está sendo desenvolvido para que todos eles sejam marcados, por isso, o descarte seria muito danoso para esse objetivo do método.

4.4 Conclusões

Este capítulo teve como objetivo desenvolver regras para fazer com que os sinais de informação estudados nos Capítulos 2 e 3 sejam utilizados para definição de quando deve ser marcado um bom momento ou não.

Vimos que cada sinal de informação tem um perfil de marcações diferente, e que, portanto, necessitam ser interpretados de maneiras distintas para realizar a detecção de bons momentos. Ficou claro que o sinal de informação baseado na energia possui uma certeza maior ao realizar marcações em seqüência, e que se pode ter a confiança de que ali muito provavelmente é uma região de bom momento, ao contrário do sinal baseado no *pitch* que fazia algumas marcações inúteis ao sistema.

Mas por outro lado, vimos que o sinal baseado no *pitch* é mais rápido ao fazer marcações do que o baseado na energia, o que é bastante útil para determinar os limites do bom momento já que o sinal baseado na energia já descobriu que tal região realmente é de bom momento.

Além disso, analisando os intervalos entre trechos marcados, foi possível verificar que muitas vezes trechos vizinhos eram um só bom momento, e assim, foi desenvolvido um método para reunir trechos vizinhos nesses casos, que reduziu drasticamente o número de trechos marcados.

Tentamos ainda, verificar se trechos de duração curta poderiam ser considerados erros do método, porém foi visto que há uma parcela substancial dos bons momentos que são curtos. Assim, o descarte não deve ser implementado. Portanto, com o desenvolvimento do módulo de decisão, temos um sistema pronto para detectar bons momentos dentro de um programa esportivo. O próximo capítulo terá o objetivo de avaliar o funcionamento do sistema em diversas situações.

Capítulo 5 - Resultados

Ao fim do desenvolvimento do método de detecção automática de bons momentos, o próximo passo é testá-lo. Para isso, foi interessante desenvolver uma ferramenta com o objetivo de facilitar a aplicação do método e ter retorno dos resultados. Após, temos que elaborar uma maneira para mensurar os resultados e tornar possível comparações entre os sinais. Os testes foram separados em três etapas. A primeira é validar o método com o sinal de áudio usado de base para todo o estudo do método. Depois, devemos testar com um sinal diferente, mas do mesmo narrador. E por fim, testar com sinais de narradores diferentes. Assim, será possível avaliar se o método é capaz de marcar satisfatoriamente os melhores momentos de uma partida de futebol em diversas situações.

Na seção 5.1 descreveremos a ferramenta desenvolvida que implementa o método de detecção de melhores momentos. A seção 5.2 criará um modelo para

avaliação dos resultados. A seção 5.3 fará avaliações iniciais em todos os jogos, e na seção 5.4 serão feitas avaliações com os limiares dos sinais de informação ajustados. A seção 5.5 tratará e avaliará os erros que o método apontou. A seção 5.6 trará a avaliação de um profissional da área de TV que utilizou o sistema. por fim, a seção 5.7 relembrará o que foi feito neste capítulo e trará comentários.

5.1 *Software* MelhoresMomentos

5.1.1 Descrição

O *software* **MelhoresMomentos** é a aplicação do método discutido neste projeto. Com ele, o usuário busca os melhores momentos de uma partida desejada. Ele foi desenvolvido em C++ com base em [7] e [8] e utilizando MFC 8.0, biblioteca do Windows referenciada em [9], IT++ 4.0.0, biblioteca para processamento de sinais referenciada em [10], que utiliza a biblioteca MKL 9.1.027 da Intel disponível em [11].

Figura 27 - Interface do MelhoresMomentos.

Com o *software*, o usuário é capaz de abrir um arquivo de vídeo, tocar e parar, selecionar um trecho, exportar tanto áudio como vídeo, e detectar os melhores momentos existentes no trecho selecionado. Na Figura 27 vemos a janela principal do aplicativo, que está descrita a seguir:

1 – *Display* de exibição do vídeo carregado;

2 – *Timeline* do vídeo, onde aparece pintado o trecho selecionado nos botões *MarkIn* e *MarkOut*;

3 – Botões de *Play* e *Stop*;

4 – Caixa que exibe *status* da ação que está sendo realizada;

5 – Botão que seleciona o ponto de entrada, em *frames*, do vídeo a sofrer a ação;

6 – Botão que seleciona o ponto de saída, em *frames*, do vídeo a sofrer a ação;

7 – Posição, em *frames*, do vídeo no *timeline*.

8 – Lista com os melhores momentos, em *frames*, encontrados pelo *software*, onde se clicar em algum, o início será posicionado no *timeline*; e

9 – Botão para limpar a lista de melhores momentos.

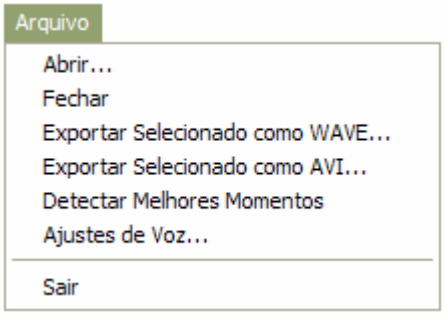

Figura 28 - Menu Arquivo do MelhoresMomentos, onde são chamadas as funcionalidades do sistema.

No menu **Arquivo** localizado abaixo do *caption* do *software* e exibido na Figura 28, estão presentes as chamadas para as funcionalidades do sistema, que estão descritas a seguir:

- **Abrir... –** Abre janela para buscar o vídeo que será carregado;
- **Fechar –** Fecha o vídeo que está carregado;
- **Exportar Selecionado como WAVE... –** Abre janela para indicar onde deve ser gravado o arquivo de áudio WAV do trecho selecionado pelo usuário;
- **Exportar Selecionado como AVI... –** Abre janela para indicar onde deve ser gravado o arquivo de vídeo AVI do trecho selecionado pelo usuário;
- **Detectar Melhores Momentos –** Inicia a busca por melhores momentos no trecho selecionado pelo usuário;
- **Ajustes de Voz... –** Abre janela mostrada na Figura 29, onde são feitos ajustes para a detecção dos melhores momentos; e
- **Sair –** Sai do sistema.

Na Figura 29 são realizados os ajustes, que mais tarde chamaremos de normalização. Há dois ajustes a serem feitos, o primeiro é o Volume, que variará o limiar do sinal de informação baseado na energia, e o segundo é o ajuste de Tom, que variará o limiar do sinal de informação baseado no *pitch*.

Figura 29 - Janela onde são realizados ajustes de Volume e Tom do víedo para a detecção de melhores momentos.

5.1.2 Organização do Código

O **MelhoresMomentos** foi desenvolvido da forma mais simples possível, porém sem deixar de ser organizada. Assim, as classes foram separadas em três camadas, como mostra a Figura 30. A primeira contém as classes responsáveis pela visualização do *software*, que são as classes **CMelhoresMomentosDlg**, responsável pela janela principal do *software*, **CVoxAdjustsDialog**, responsável pela janela de ajustes de volume e tom, e, por fim, as classes da biblioteca **MFC**, responsáveis pelos componentes gráficos, que deriva da **CwinApp**, classe base do Windows para aplicativos.

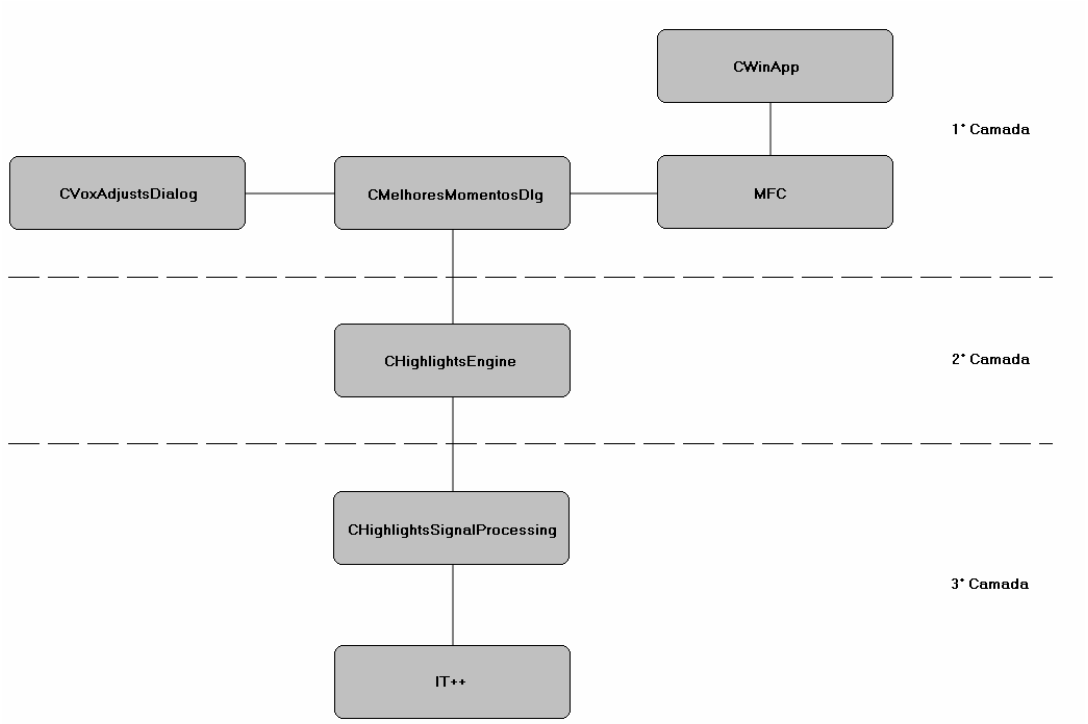

Figura 30 - Diagrama de classes do MelhoresMomentos.

A segunda camada contém a *thread* que roda as funcionalidades do sistema, tais como exportar áudio e vídeo, e detectar os melhores momentos do trecho selecionado. A **CHighlightsEngine** é uma *thread* para que as janelas do *software* não fiquem presas enquanto qualquer processamento é feito, já que eles não são instantâneos. Por exemplo, a detecção de melhores momentos de 45 minutos do jogo leva cerca de 3 minutos em um computador com um processador Intel Pentium Dual

Core 3.06 GHz. A **CHighlightsEngine** envia *status* do processamento constantemente para a primeira camada.

A terceira camada é responsável pelos cálculos necessários às funcionalidades executadas no **CHighlightsEngine**. A classe **CHighlightsSignalProcessing** é quem aplica todos os procedimentos descritos neste projeto, e para isso utiliza as classes de processamento de sinais da biblioteca **IT++**.

5.2 Modelo de Avaliação

Com o *software* **MelhoresMomentos** já temos como testar o método, assim, podemos avaliar os resultados obtidos de duas formas. A primeira é verificar quantos dos trechos que o método marcou realmente são bons momentos, que chamaremos aqui de Percentual de Trechos Marcados Corretamente (**%TMC**). E a outra é verificar quantos bons momentos que realmente existem no jogo foram marcados pelo método, que chamaremos de Percentual de Bons Momentos Marcados (**%BMM**).

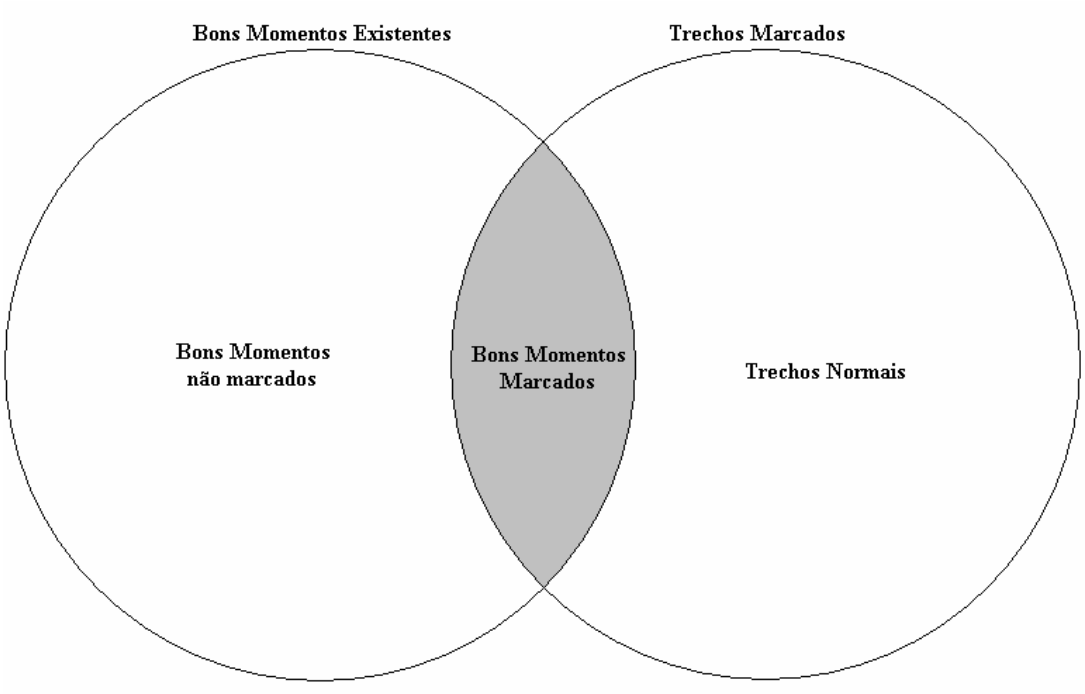

Figura 31 - Diagrama ilustrando os parâmetros que servirão para visualização dos resultados.

O diagrama da Figura 31 ilustra visualmente o que os parâmetros representam dentro dos resultados obtidos. Vale lembrar que o **%BMM** é o objetivo principal do

método, visando, assim, marcar todos os bons momentos da partida. Enquanto que o **%TMC** é desejável que tenha alto índice, porém não é essencial. Pois como vimos nos capítulos anteriores é preferível o método marcar lances normais a mais do que deixar de marcar bons momentos. Um **%TMC** alto facilitaria em uma etapa subseqüente de validação da marcação.

Além disso, podemos ao fim, tentar entender os motivos pelo qual o método marcou trechos erroneamente, para isso basta fazermos uma análise estatística do provável motivo do erro observado.

5.3 Avaliação Inicial

De início, devemos testar o método no Sinal I, que foi usado de referência para o desenvolvimento do projeto. Na tabela 4 podemos observar que apenas 58,3% dos trechos marcados pelo sistema realmente eram bons momentos, enquanto que 100% dos bons momentos existentes no jogo apareceram entre os trechos marcados. Estes resultados são esperados já que o método foi desenvolvido sobre este sinal de áudio e vislumbrava marcar todos os bons momentos, mesmo que trechos irrelevantes aparecessem. Dessa maneira, a marcação dos bons momentos passaria de automática para semi-automática, ou seja, seria necessário um usuário para descartar os trechos indesejados.

Além do primeiro tempo, testamos o método no Sinal II, que é o segundo tempo da mesma partida do Sinal I, e os resultados foram similares, sendo que este ainda apresentou o **%TMC** pouco melhor do que o trecho original.

Tabela 4 - Resultados iniciais da avaliação.

Sinal	Narrador	%TMC	%BMM
Sinal I	Narrador I	58,3	100
Sinal II	Narrador I	65,2	100
Sinal III	Narrador I	44,2	100
Sinal IV	Narrador II	87,5	43,8
Sinal V	Narrador II	0	0
Sinal VI	Narrador III	10,5	50

O próximo passo foi testar outra partida que fosse narrada pelo Narrador I, o Sinal III. Dessa maneira, testamos um sinal diferente, com a captação de áudio diferente, em um ambiente diferente, mas com as mesmas características de voz utilizadas no desenvolvimento do método. Os resultados mais uma vez foram satisfatórios, pois, apesar do **%TMC** ter diminuído, acarretando um aumento de trechos indesejados marcados, o **%BMM** continuou atingindo 100%, o que é o real objetivo do método.

Por fim, validamos o método com partidas que além da captação diferente, tenham características de voz do narrador diferentes. Os Sinais IV,V e VI narradas pelo Narrador II e III, respectivamente, apontaram resultados muito ruins, como indicado na Tabela 4. O Sinal IV teve **%TMC** alto, mas **%BMM** foi baixo. O Sinal V não teve nenhum bom momento apontado. O Sinal VI foi o que obteve o pior resultado, pois tanto **%TMC** quanto **%BMM** foram baixos. Os resultados ruins podem ser devidos aos limiares que não eram ideais para tais características de voz.

5.4 Avaliação Normalizada

Notamos que os resultados anteriores de partidas com narradores diferentes do Narrador I foram muito ruins devido aos limiares de energia e *pitch* estarem inadequados às características de voz. Assim, tentaremos normalizar os resultados, ou seja, fazer com que **%BMM** atinja 100% e assim podermos visualizar melhor a diferença de acerto do método entre os narradores. Para isso, os limiares foram variados, ajustados manualmente, até encontrar uma situação em que todos os bons momentos existentes fossem marcados.

A Tabela 5 apresenta os resultados após a normalização manual dos limiares, onde é possível ver que o **%BMM** atingiu 100% para todos os jogos, como desejávamos, mas por outro lado o **%TMC** diminui bruscamente em todas as partidas. É fácil reparar que os percentuais de trechos marcados corretamente ficaram relativamente próximos nos jogos que tinham o mesmo narrador. Também podemos ver que a partida narrada pelo Narrador III apresentou um **%TMC** muito baixo, ou seja, quase todos os trechos marcados não eram bons momentos.

Podemos reparar que a normalização fez com que **%BMM** chegasse ao desejado, porém, para que isso se tornasse possível, mais uma vez, houve a necessidade de um usuário que realizasse os ajustes de limiares para que o método incluísse o maior número de bons momentos existentes possível.

Tabela 5 - Resultados normalizados.

Sinal	Narrador	%TMC	%BMM
Sinal IV	Narrador I	35,2	100
Sinal V	Narrador I	23,4	100
Sinal VI	Narrador I	7,3	100

A normalização resolveu a identificação de todos bons momentos existentes nas partidas. Uma maneira de facilitar a normalização de limiares seria possibilitar que o usuário escolhesse um trecho de bom momento e outro normal para que o método calculasse os limiares automaticamente. Outra maneira seria montar um banco de narradores com seus limiares pré-definidos.

Tabela 6 - Quantidade de bons momentos que tiveram seus limites marcados satisfatoriamente pelo método.

Sinal	Narrador	%BMS
Sinal I	Narrador I	64,3
Sinal II	Narrador I	73,3
Sinal III	Narrador I	73,7
Sinal IV	Narrador II	68
Sinal V	Narrador II	76,5
Sinal VI	Narrador III	66,6

Agora, temos de avaliar se os inícios indicados realmente podem ser aceitos. A única forma de avaliar isso é validar manualmente se o bom momento já estava acontecendo na posição em que o método indicou. A Tabela 6 indica qual o percentual de bons momentos que teve seu início satisfatório, **%BMS**. Observando os bons momentos que não tiveram o início marcado corretamente, claramente o narrador demorou a aplicar emoção a voz e quando o fez, aumentou o volume rapidamente. Repare que neste aspecto todas as partidas tiveram resultados semelhantes, o que nos faz pensar que o acerto ou erro nesse quesito independe do

locutor. Isto nos leva a crer que o sistema que aplicar o método tem que permitir que o usuário redefina os limites do bom momento.

5.5 Estudo dos Erros

Sabemos que o método de detecção desenvolvido vai além de um marcador de bons momentos, na realidade trata-se de um marcador de trechos em que há emoção intensa na voz. Assim, trechos em que o narrador aplica emoção, mas que são descorrelacionados com a partida ou não se caracterizam como um bom momento, o método irá marcar também, tais como anúncios, *replays*, inícios e términos da partida e momentos pós-gol. Dessa forma, outra análise que podemos realizar é tentar entender porque o método apontou bom momento em um trecho quando ele não era.

Tabela 7 - Distribuição dos erros por motivos e partidas.

Sinal	Narrador	Emoção (%)	Outra Pessoa Falando (%)	Sem Motivo Aparente (%)
Sinal I	Narrador I	60	40	0
Sinal II	Narrador I	62,5	25	12,5
Sinal III	Narrador I	78	17,4	4,6
Sinal IV	Narrador II	53,3	16,7	30
Sinal V	Narrador II	41,7	16,6	41,7
Sinal VI	Narrador III	32	8	60

Podemos, então, separar os erros em três classes. Os momentos em que o narrador aplica emoção na voz, mas que não se caracterizam bons momentos, os momentos em que há outra pessoa falando, tais como comentarista, repórter, jogadores e árbitros, e os momentos em que não há um motivo aparente para o método marcá-lo. Nesse contexto, os dois últimos casos são considerados erros de detecção, pois não há bons momentos narrados nesses trechos, e, portanto, não deveriam ser marcados. Já no primeiro caso há emoção na voz do narrador, ou seja, o método acertou ao marcar, porém, a nossa aplicação visa marcar somente bons momentos, o que esses trechos, apesar de terem emoção, não são. A Tabela 7 demonstra que uma grande parcela dos erros do detector de bons momentos não são erros do detector de emoção. Dessa forma, a Tabela 8 ilustra os resultados, caso o

detector de emoção fosse suficiente para nossa aplicação, onde verificamos que o método atinge valores satisfatórios, até nos jogos narrados pelo Narrador II.

Tabela 8 - Resultados normalizados considerando erros com emoção.

Sinal	Narrador	%TMC	%BMM
Sinal I	Narrador I	83,3	100
Sinal II	Narrador I	86,9	100
Sinal III	Narrador I	86	100
Sinal IV	Narrador II	70,1	100
Sinal V	Narrador II	66,7	100
Sinal VI	Narrador III	39,3	100

Além disso, podemos fazer uma última análise baseada na duração dos trechos marcados. No final do Capítulo 4, vimos que o descarte de trechos curtos seria maléfico para o método, pois ele atingia trechos de bons momentos. Porém, podemos ver se isso será realidade também para os demais jogos e narradores.

Tabela 9 - Percentual de trechos curtos marcados e percentual de trechos curtos que são bons momentos.

Sinal	Narrador	%TC	%TCBM
Sinal I	Narrador I	45,8	36,4
Sinal II	Narrador I	43,5	60
Sinal III	Narrador I	46,5	30
Sinal IV	Narrador II	57,9	36,4
Sinal V	Narrador II	76,5	23,1
Sinal VI	Narrador III	55,4	0

Na Tabela 7, vemos que estes trechos são freqüentes nas marcações realizadas, como aponta **%TC**, percentual de trechos curtos. Mas vemos também que em quase todos os jogos, havia bons momentos dentre os trechos curtos, como indica **%BMTC** (percentual dos trechos curtos que são bons momentos). No Sinal VI narrado pelo Narrador III, vemos que não há nenhum bom momento entre os trechos curtos, e mais da metade das marcações são curtas. Portanto, para esse caso, o descarte seria proveitoso, mas para todos os outros, perderíamos trechos desejáveis.

5.6 Avaliação Profissional

Na avaliação do Coordenador de Projetos de Sistemas de TV da TV Globo, Daniel Monteiro de Barros, a idéia de construir um sistema de marcação automática de bons momentos seria inovadora, pois o processo de marcação manual, edição e composição de um clipe de melhores momentos depende de diversos sistemas e pelo menos um operador por partida. Em sua visão, seria interessante o operador escolher os bons momentos para formar um clipe e os exportar como vídeo dentro do mesmo sistema. Daniel, que utilizou o sistema, diz este ainda não tem condições de entrar em operação, pois o operador tem que realizar ajustes grosseiros nos limiares de volume e tom, o que pode vir a ser trabalhoso.

5.7 Conclusões

A primeira tarefa deste capítulo foi mostrar a ferramenta que foi criada para aplicar o método de marcação automática de bons momentos desenvolvido neste projeto, chamado de **MelhoresMomentos**. O funcionamento e as funcionalidades do *software* foram descritos, além da organização do código. Com esta ferramenta, tornou-se mais fácil realizar os testes posteriores do método.

Além disso, foi criado um modelo de avaliação que tinha como objetivo validar tanto a eficiência do método ao marcar somente pontos corretos, quanto ao marcar todos os pontos desejados da partida de futebol. Com ele, foi possível verificar que nos jogos, que tiveram a mesma narração da partida utilizada de base para o desenvolvimento, incluíram em suas marcações todos os trechos relevantes, mas também incluíram alguns trechos não relevantes. Isso nos levou a concluir que o método não deve ser automático, mas semi-automático, pois assim, haveria um usuário que pudesse descartar os trechos não relevantes que aparecessem.

Após, vimos que as partidas com narradores diferentes não conseguiam marcar todos os bons momentos existentes, o que era bastante crítico para o método. Dessa maneira, tivemos que criar controles para que o usuário ajustasse os limiares de energia e *pitch* e, assim, todos os bons momentos pudessem ser marcados.

Foi possível ver que os erros para detectar bons momentos contêm alguns acertos para detectar trechos de emoção, e que o método é capaz de marcar todos os trechos que tenham emoção na voz do narrador. Isso nos mostrou que se utilizarmos um método semi-automático, ajustando limiares e descartando trechos indesejados, mesmo com narradores diferentes o método conseguirá marcar todos os bons momentos existentes na partida, o que é o principal objetivo deste Projeto Final.

Por fim, vimos que independentemente do narrador, quando o método encontra o trecho de bom momento, em mais da metade das vezes ele consegue determinar início e fim adequados. Se o sistema que implementa o método permitir a remarcação dos limites, isso seria resolvido pelo usuário.

Capítulo 6 - Conclusão

Neste projeto desenvolvemos um método para marcação automática de bons momentos em uma partida de futebol. Neste capítulo relembraremos o que foi feito, enfatizaremos as contribuições do projeto, e, por fim, indicaremos novos caminhos.

6.1 Contribuições

O método desenvolvido neste projeto tem o claro objetivo de reduzir os custos operacionais da rotina de marcação de bons momentos em transmissões televisivas. De início, foi idealizado que somente ao aplicar o método em qualquer programa esportivo, fossem retornados todos os seus bons momentos. Porém, foi visto que o método automático não seria possível, então se optou pelo semi-automático, onde haveria um usuário que faria ajustes e descartaria trechos marcados erroneamente. Mesmo assim, o método traria melhorias às transmissões televisivas,

já que no Sinal I, por exemplo, em 47 minutos e 10 segundos de partida, o método retornou 4 minutos e 25 segundos para serem revisados. Há uma economia de tempo considerável, pois o operador não teria mais necessidade de assistir ao jogo.

Além disso, houve o desenvolvimento de um *software* que implementa o método estudado, que apesar de ter um desempenho computacional abaixo do esperado, pode ser utilizado.

Na opinião de um profissional de TV, quando o sistema generalizar o método para diversos locutores, somente um operador poderá ser responsável por gerar clipes de melhores momentos de partidas que ocorrem simultaneamente, além de todo o processo poder ser feito em um só sistema.

Entretanto, a maior contribuição deste projeto é dar início aos estudos de voz relacionados à identificação e marcação automática de eventos em programas esportivos transmitidos pela TV. Ter sistemas com tais objetivos funcionando em condições de entrar em operação em uma empresa de TV, traria aperfeiçoamentos consideráveis ao processo de melhores momentos de uma transmissão televisiva.

6.2 Retrospectiva

O Capítulo 1 forneceu a motivação para o desenvolvimento do projeto e foi responsável por uma idéia geral de como funcionaria o método, onde foi visto que dois caminhos poderiam ser seguidos, da energia e freqüência fundamental da voz.

O Capítulo 2 estudou diversas maneiras para gerar o sinal de informação baseado na energia, além de analisá-los para chegar ao limiar de energia necessário para marcar bons momentos. Chegou-se a conclusão que a melhor maneira seria utilizar o janelamento superposto de 1000 ms, e utilizar limiar de $E = 0.04$.

O Capítulo 3 realizou um estudo similar ao do Capítulo 2, porém para a freqüência fundamental da voz, a freqüência de *pitch*. No fim, foi visto que uma janela com média de *pitch* maior que o limiar de 200 Hz seria um bom momento.

O Capítulo 4 foi responsável pelo módulo de decisão do método, onde foi implementada a inteligência que observa os dois sinais de informação e decide se é bom momento ou não. Nos estudos, foi visto que o sinal de informação baseado na energia será o responsável por determinar a região onde se encontra o bom momento, enquanto que o baseado no *pitch* ajudará na determinação de início e fim. O módulo de decisão ficou responsável também por unir bons momentos muito próximos.

O Capítulo 5 começou descrevendo a ferramenta desenvolvida que implementa o método de marcação de melhores momentos. Além disso, detalhou como seriam feitos os testes de validação e avaliação do método e o testou. Os resultados mostraram que o método funciona bem para o mesmo narrador que foi utilizado no desenvolvimento do projeto, mas para os demais se mostrou pouco eficiente. Assim, foram feitos ajustes dos limiares para atender ao objetivo do método que é marcar todos os bons momentos existentes, o que ocasionou o aumento de trechos indesejados no resultado. Depois, foi visto nos erros apontados que a grande maioria era de momentos em que o narrador aplicava emoção, e, assim, se tratavam de erros para a aplicação em bons momentos. Mas não do método de detecção de emoção, que se mostrou bastante eficaz na marcação de trechos com emoção simplesmente. E, ao fim do capítulo, chegou-se a conclusão de que o método deve ser semi-automático, e não automático como era pensado no início, tanto para ajustar os limiares, quanto para descartar trechos indesejados e redefinir início e fim.

6.3 Propostas para Trabalhos Futuros

Este Projeto Final, como já foi dito, é somente o início de um estudo relacionado a processamento de voz para identificação e marcação de eventos em transmissões televisivas. Desse modo, certamente novos estudos podem ser feitos, novas abordagens podem ser agregadas e melhorias efetuadas. A seguir, uma listagem com algumas propostas:

- Adicionar funcionalidades ao **MelhoresMomentos**, tais como permitir a redefinição dos limites do bom momento, e permitir a criação de clipes com os trechos selecionados pelo usuário;

- Otimizar o **MelhoresMomentos**, tais como paralelizar o processamento, e implementar cascateamento de classificadores;
- Implementar melhorias na detecção de *pitch*, tais como a detecção de trechos sonoros e surdos;
- Estudo de uma forma de normalizar automaticamente os sinais de áudio, para que o método independa de narrador e captação;
- Aperfeiçoamento do módulo de decisão, tal como estudo dos trechos curtos para o possível descarte;
- Utilizar outras abordagens do estudo de voz, a fim de obter mais um meio de determinar bons momentos, tais como reconhecimento de palavras específicas;
- Estudo da aplicação do método para outros esportes além do futebol.

Referências bibliográficas:

[1]. D. Rocchesso, Introduction to Sound Processing, [http://www.mondo-estremo.com], Mondo Estremo, 20/03/2003.

[2]. P. S. R. Diniz, E. A. B. Da Silva, S. L. Netto, Processamento Digital de Sinais – Projeto e Análise de Sinais, Bookman Editora, 2004.

[3]. P. Z. Peebles, Probability, Random Variables, and Random Signal Principles, McGraw-Hill, 2001.

[4]. T. Tolonen, M. Karjalainen, A Computationally Efficient Multipitch Analysis Model, IEEE Trans. Speech Audio Processing, 8(6), 708-716, Nov. 2000.

[5]. D. Coulter, Digital Audio Processing, CMP Books, 2000.

[6]. J. H. Deller, J. R. Proakis, J. G. Hansen, Discrete-Time Processing of Speech Signals, Prentice Hall, 1987.

[7]. P. M. Embree, D. Danieli, Algorithms for Digital Signal Processing.

[8]. N. M. Josuttis, The C++ Standard Library – A Tutorial and Reference, Addison-Wesley, Nov. 2006.

[9]. Microsoft Development Network, [http://www.msdn.com].

[10]. IT++ 4.0.0, [http://itpp.sourceforge.net/], 14/10/2007.

[11]. Intel Math Kernel Library 9.1.027, [http://www.intel.com/cd/software/products/asmo-na/eng/307757.htm]