Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Marcação Automática de Eventos Usando Sinal de Áudio em Transmissões Esportivas de TV

Luiz G. L. B. M. de Vasconcelos[1], Sergio L. Netto[2], Luiz W. P. Biscainho[2] e Charles B. do Prado[1]
[1]Departamento de Pesquisa e Desenvolvimento, TV Globo
Rio de Janeiro, RJ, 22460-000, Brasil
[2]Programa de Engenharia Elétrica/COPPE, DEL/Poli, Universidade Federal do Rio de Janeiro
CP 68504, Rio de Janeiro, RJ, 21941-972, Brasil
luiz.vasconcelos@tvglobo.com.br, sergioln@lps.ufrj.br, wagner@lps.ufrj.br, charles.prado@tvglobo.com.br

RESUMO

Este artigo descreve um método para localizar os melhores momentos da transmissão de um jogo de futebol a partir do áudio, com base na energia e na freqüência fundamental da voz do narrador. Para isso, implementou-se um aplicativo com interface gráfica que permite classificar o sinal de forma rápida e prática. O sistema mostrou-se capaz de identificar 100% dos momentos de interesse para o mesmo narrador utilizado no treinamento, ao custo de uma taxa de falsa identificação em torno de 50%. O processo de seleção comprime o sinal de vídeo em cerca de 90% para uma posterior classificação semi-automática.

0 INTRODUÇÃO

Cada vez mais, em nossa sociedade, aumenta a demanda por entretenimento, tal como acesso à Internet, peças teatrais e cinematográficas, *shows* de música, prática de esportes e viagens. Nesse contexto, se inserem as emissoras de TV, que, além de informar, também têm o objetivo de entreter. Uma parcela substancial do entretenimento televisivo é a transmissão de programas esportivos, tais como partidas de futebol, e ainda a posterior exibição de eventos específicos, tais como gols, pênaltis, oportunidades de gol etc. O interesse por transmissões esportivas é tão grande que há canais com programação dedicada a elas, que também exibem eventos específicos ocorridos em outras transmissões e mesmo programas secundários que noticiam apenas esses eventos. Preparar esse tipo de programação requer um enorme consumo de tempo e esforço para cobrir todas as transmissões, já que se requer uma seleção bastante criteriosa dos eventos de interesse. Atualmente, é necessário um operador acompanhando cada transmissão e marcando os eventos específicos para posterior recuperação, o que torna interessante o desenvolvimento de tecnologias que automatizem ou simplifiquem este processo.

A princípio, pode-se considerar o processamento das imagens para detectar padrões em transmissões televisivas. Porém, retirar informações tão específicas do sinal de vídeo seria bastante complexo, pois cada esporte tem caracterísitcas visuais próprias, além do fato de este tipo de processamento envolver um volume muito grande de dados. Sendo assim, já há trabalhos [1] que analisam o áudio para encontrar trechos desejados de um sinal de vídeo.

Analisando a Figura 1, tem-se que o modelo de produção da voz humana considera um sinal de excitação processado por um filtro que modela o trato vocal. A excitação,

proveniente dos pulmões, caracteriza um aspecto da sonoridade associado à vibração (trecho sonoro) ou não (trecho surdo) das cordas vocais. Para todos os efeitos práticos, em processamento de voz, a freqüência de vibração das cordas vocais é denominada de *pitch*. O sistema aqui proposto de classificação de "bons momentos" se baseia nas informações de energia e de *pitch* do sinal de voz. De modo geral, esses dois parâmetros se elevam de forma significativa durante os eventos de interesse.

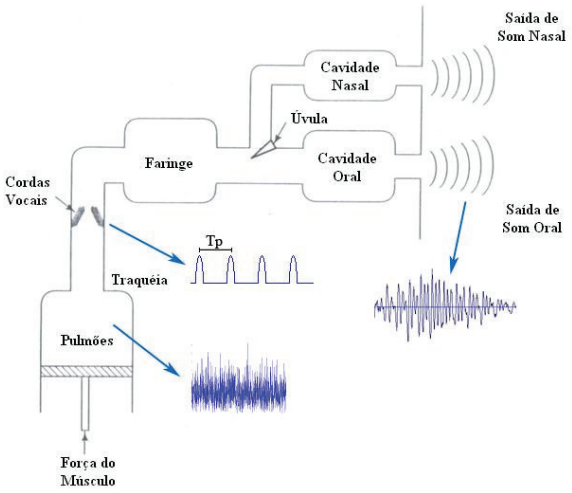

Figura 1 Representação em blocos do processo de geração da voz humana [2].

Nesse contexto, é feita uma análise da evolução destes dois aspectos ao longo dos trechos de interesse para um dado sinal de treinamento. Posteriormente, um módulo de decisão utiliza estes sinais para classificar o trecho em questão como sendo de interesse ou não. Uma etapa final é responsável por agrupar trechos muito próximos, que representariam o mesmo evento, e também verificar a correta duração (início e fim) dos eventos previamente selecionados. O sistema incorpora uma *interface* gráfica para facilitar o acompanhamento do processo por parte do usuário, que pode ainda fazer pequenos ajustes para melhorar o processo de classificação.

Para apresentação completa do sistema, este artigo obedece a seguinte estruturação: A Seção 1 inclui uma descrição do desenvolvimento do sistema e de seu funcionamento geral. Na Seção 2, é descrita a ferramenta gráfica desenvolvida para a aplicação em questão, destacando suas principais funcionalidades no processo de classificação e edição do sinal de vídeo resultante. Na Seção 3, é caracterizado o desempenho do sistema em termos da capacidade de detecção dos "bons momentos"; são considerados sinais do mesmo narrador usado no desenvolvimento do sistema e também de outros narradores. Por fim, na Seção 4, são apresentadas as conclusões do trabalho, ressaltando-se suas principais contribuições.

1 DESENVOLVIMENTO DO SISTEMA

Por se tratar de uma aplicação bastante particular de processamento de voz, esta seção descreve o desenvolvimento do método proposto. Inicialmente, o sistema foi modelado usando-se um único sinal da base de dados, para ao final ser generalizado para outros sinais (do mesmo narrador ou não).

1.1 Base de Dados

Os sinais que compõem a base de dados usada no desenvolvimento e teste do sistema foram cedidos pela TV Globo, e são descritos na Tabela 1. Trata-se de sinais digitais de vídeo com áudio *embedded*, onde o *stream* de áudio foi amostrado à taxa de 48 kHz com 16 bits por amostra em dois canais, sendo o esquerdo referente à narração e o direito, ao ambiente. O sistema proposto é baseado apenas no sinal de narração que possui um mínimo nível de ruído ambiente. O Sinal I foi utilizado para o desenvolvimento do método. Os demais sinais, do mesmo narrador que o Sinal I ou não, foram utilizados na etapa de testes de desempenho. A Tabela 2 descreve a quantidade de "bons momentos" em cada sinal da Tabela 1. Estes valores foram obtidos de forma tradicional, isto é, determinados visualmente por um operador humano.

Tabela 1 Sinais que compõem a base de dados usada no desenvolvimento e teste do sistema.

Sinal	*Partida*	*Narrador*	*Nome*
Sinal I	Vasco x Flamengo 1°T	Narrador I	Eduardo Moreno
Sinal II	Vasco x Flamengo 2°T	Narrador I	Eduardo Moreno
Sinal III	Chivas x San Jose	Narrador I	Eduardo Moreno
Sinal IV	Botafogo x Vasco	Narrador II	Galvão Bueno
Sinal V	Brasil x Chile	Narrador II	Galvão Bueno
Sinal VI	Boca Jrs. x Grêmio	Narrador III	Cléber Machado

Tabela 2 Número de "bons momentos" para cada sinal da base de dados descrita na Tabela 1.

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

1.2 Energia do Sinal de Voz

A alta energia de um sinal de voz pode indicar se o trecho de vídeo correspondente é de interesse ou não. Para se minimizar a quantidade de dados processados, divide-se o sinal de voz $x(n)$ em blocos de N amostras e determina-se a energia E do bloco por

$$E = \sum_{n=1}^{N} x^2(n) \cdot \qquad (1)$$

O valor adequado para N pode ser determinado de forma experimental para a aplicação em questão. Valores pequenos geram um número excessivo de blocos, o que aumenta o custo computacional do método de classificação; por outro lado, valores excessivos para N acarretam a não-detecção de alguns trechos de interesse do sinal de vídeo. A Figura 2 ilustra dois exemplos de fala intensa com durações bastante distintas: 200 ms e 2000 ms. Com base nisso, foram considerados blocos de durações 250ms, 500ms e 1000ms para se verificar qual destes valores gera um sinal de energia que melhor destaca os momentos de interesse.

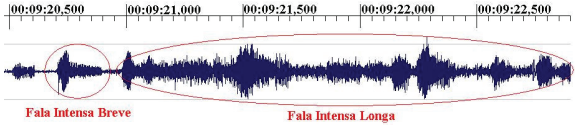

Figura 2 Exemplo de sinal de voz com trechos de interesse de diferentes durações (cerca de 200 ms e 2 s).

A segmentação do sinal de voz pode ser feita ainda usando-se blocos superpostos ou não, como indicado na Figura 3. A não-superposição ocorre quando o deslocamento M do bloco é maior ou igual ao número de amostras N que o compõem. Já a superposição resulta da condição $M < N$. Com o deslocamento superposto, o sistema carrega mais informação a respeito das variações de energia do sinal. Por outro lado, o deslocamento não-superposto é muito mais leve computacionalmente. Porém, para $M = 1$ a questão computacional do cálculo da energia do bloco é facilmente resolvida com a aplicação do algoritmo de *buffer* circular [3,4]. Nesse caso, a energia do bloco atual é determinada pela energia do bloco anterior adicionada à energia da amostra atual $x(n)$ e subtraída da energia da amostra $x(n\text{-}M)$. Para se determinar o deslocamento que melhor realça os "bons momentos", serão considerados os casos $M = 1$ e $M = N$.

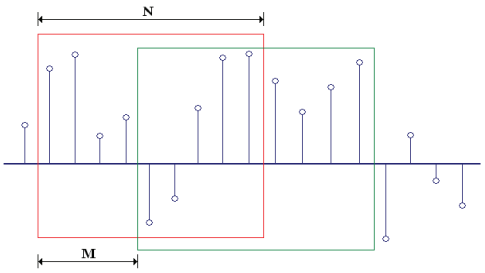

Figura 3 Deslocamento do bloco no domínio do tempo.

Desta forma, seis variações de segmentação foram testadas na classificação do Sinal I: com durações de 250, 500 e 1000 ms, e ainda $M = 1$ e $M = N$ para cada caso anterior. Para cada variação, foi feito um histograma das distribuições da energia dos blocos associados aos eventos de interesse ou não. De modo geral, em todos os casos foi possível observar uma boa separação dos histogramas associados a cada tipo de bloco, como é visto na Figura 4 para a duração de 250 ms e $M = 1$. As Figuras 5 ($M = 1$) e 6 ($M = N$) mostram a taxa de classificação correta de cada tipo de bloco (evento de interesse ou não) em função do limiar de decisão escolhido para a energia do bloco, para as seis variações acima descritas. Destas figuras, conclui-se que todas as variações possuem desempenho semelhante, com uma pequena vantagem em termos de taxa de classificação para o caso $M = 1$ com duração de 1000 ms. Para este caso, privilegiando-se a identificação correta dos "bons momentos" em detrimento de uma identificação incorreta de alguns momentos normais, pode-se estipular o limiar de energia como sendo de 0,04. Naturalmente, este valor é altamente dependente do nível de gravação do sinal de entrada, mas uma normalização apropriada pode ser feita para torná-lo de uso mais geral.

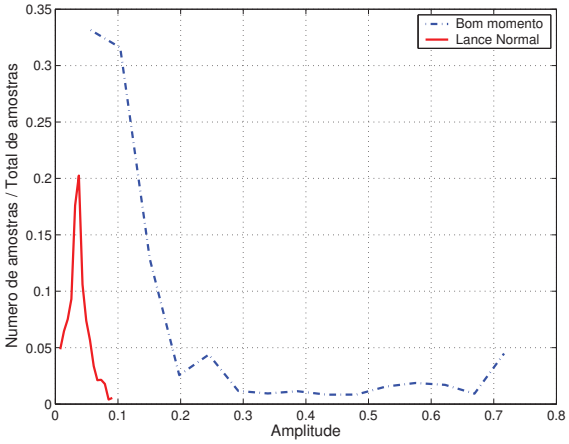

Figura 4 Distribuição estatística do sinal de energia calculada com duração de 250 ms e janela superposta.

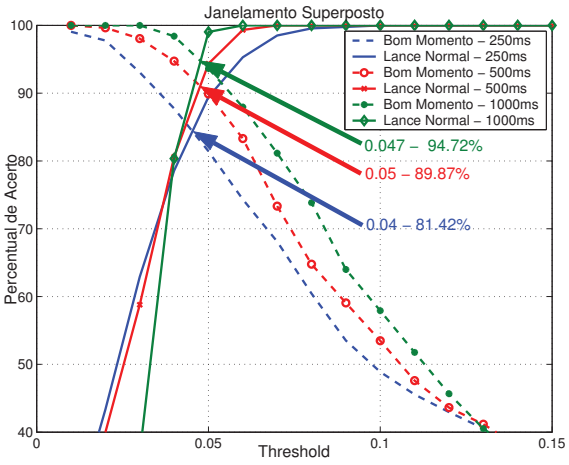

Figura 5 Taxas de acerto de classificação em função do limiar de energia para $M = 1$ e duração de 250, 500 e 1000 ms.

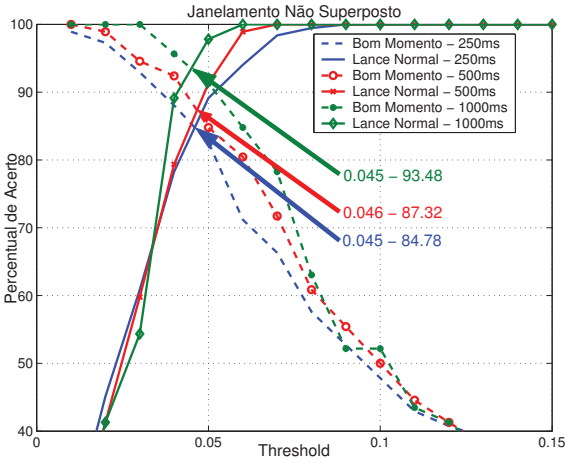

Figura 6 Taxas de acerto de classificação em função do limiar de energia para $M = N$ e duração de 250, 500 e 1000 ms.

1.3 Pitch do Sinal de Voz

O período de *pitch* da voz é determinado pelos movimentos quase periódicos das cordas vocais na faringe, e é o inverso da freqüência fundamental da voz percebida pelo sistema auditivo humano [2].

Uma maneira de se extrair a freqüência fundamental da voz é determinar sua periodicidade a partir da função de autocorrelação [5,6]:

$$R_{xx}(\tau) = \sum_{n} x_n x_{n-\tau} \cdot \qquad (2)$$

A Figura 7 ilustra o aspecto da função de autocorrelação para trechos do sinal de voz associados a momentos de interesse ou não. A partir desta figura, é possível perceber que os picos mais proeminentes da autocorrelação, que determinam o período de *pitch* do trecho de voz correspondente, ficam mais próximos entre si nos caso de um "bom momento".

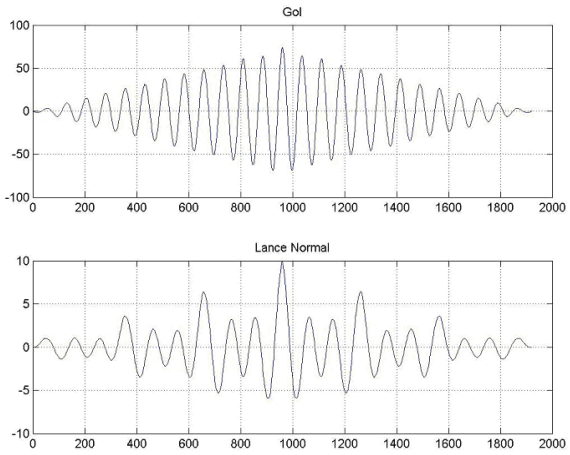

Figura 7 Autocorrelação do sinal de voz para um trecho de gol e outro de lance normal.

O cálculo da autocorrelação pode ser feito no domínio da freqüência a partir da relação [7]:

$$R_{xx}(\tau) = IDFT\{|\,DFT[x(n)]\,|^2\}. \qquad (3)$$

A freqüência fundamental da voz masculina está em geral em torno de 150 Hz, ou ao menos acima dos 80 Hz. Assim, um período de *pitch* será no máximo de 12,5 ms. A fim de realizar o cálculo do *pitch* pela autocorrelação de forma precisa, é interessante ter pelo menos três ciclos no sinal de voz. Para forçar uma margem de segurança, foram utilizados blocos de 40 ms, correndo-se um pequeno risco de modelar pequenas variações de *pitch* dentro de um único bloco. Para evitar interferências provenientes de outras fontes, antes de qualquer cálculo foi feita uma filtragem passa-baixas limitando a banda do sinal de voz em 1 kHz. Para evitar cálculos desnecessários, foi feita uma detecção de silêncio usando-se um limiar de 0,1 para a energia de cada bloco de 40 ms do sinal de voz em questão. Este valor limite foi determinado a partir de uma análise estatística da energia para os blocos de silêncio ou não em todo o Sinal I, como ilustrado na Figura 8.

A Figura 9 exibe os histogramas do valor de *pitch* dos segmentos marcados como "bons momentos" ou não. É fácil ver que o cruzamento das distribuições se encontra na freqüência de *pitch* igual a 225 Hz. Porém, mais uma vez, por ser mais importante classificar corretamente todos os eventos de interesse, é desejável utilizar um limiar de classificação ligeiramente menor. Usando-se 200 Hz, 87,5% dos blocos de "bom momento" e 3,5% dos lances normais estão sendo marcados. Estes índices podem ser considerados satisfatórios, pois os demais blocos de interesse podem ser identificados pela continuidade do sinal.

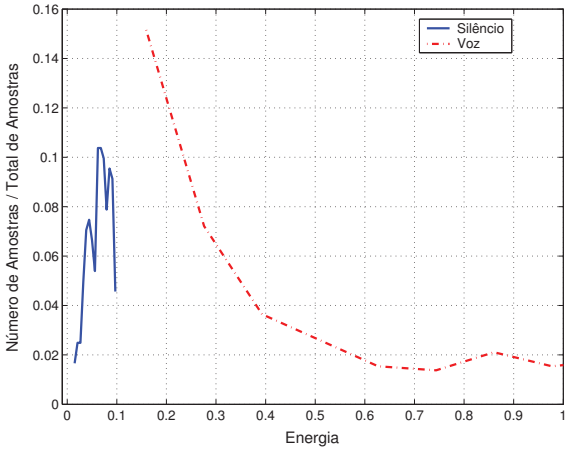

Figura 8 Histogramas de energia para trechos de silêncio e voz do Sinal I.

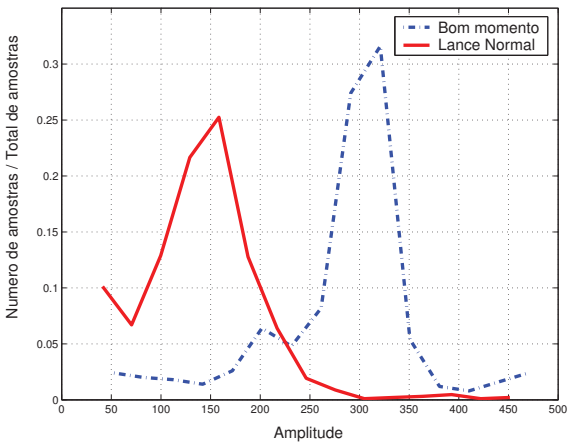

Figura 9 Distribuição do pitch em blocos de eventos de interesse ou não para o Sinal I.

Os limiares aqui encontrados para o valor de *pitch* devem ser válidos para quaisquer sinais do mesmo narrador usado no Sinal I. Para outros narradores, uma análise similar deve ser feita *a priori*, ou ainda de forma automática a partir de um trecho curto do sinal.

1.4 Módulo de Decisão

Os limiares de energia e *pitch* determinados anteriormente servem para um primeiro nível de classificação de um dado bloco como sendo de "bom momento" ou não. A Figura 10 ilustra um exemplo de marcações de um trecho do sinal de voz, onde é possível constatar que a marcação bloco-a-bloco funcionou de forma semelhante para as duas características (energia e *pitch*). De modo geral, para os limiares pré-determinados acima, observa-se que o sinal de energia foi mais conservador no sentido de que suas marcações estavam quase sempre corretas, porém demoravam mais a

identificar um trecho de interesse. Então, foi utilizado um algoritmo que buscasse pelas regiões de bom momento através da energia, para posteriormente confirmar e definir seus limites a partir do sinal de *pitch*.

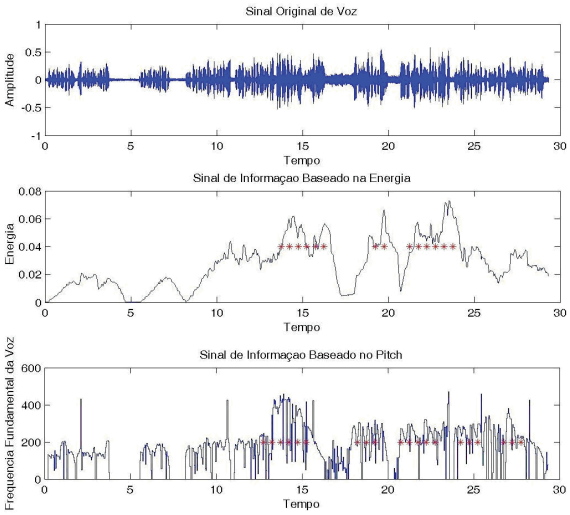

Figura 10 Sinais de informação com marcações instantâneas.

Foi feito ainda um estudo de quantas marcações de energia em seqüência são necessárias para se caracterizar de forma efetiva um "bom momento" no Sinal I. O gráfico da Figura 11 mostra que quanto maior a exigência no número mínimo de marcações em seqüência, menor é o percentual de "bons momentos" identificados. Para garantir a identificação de todos os trechos de interesse e eliminar alguns trechos erroneamente identificados anteriormente, foi então adotada a exigência mínima de três marcações em seqüência para classificar um trecho como "bom momento".

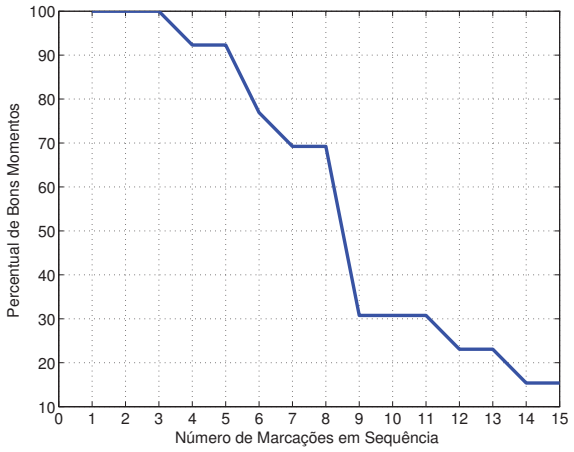

Figura 11 Percentual de "bons momentos" identificados em função do número mínimo de marcações em seqüência.

A Figura 12 é o resultado da aplicação do algoritmo no exemplo da Figura 10. É possível notar que o trecho que foi marcado pela energia com apenas duas marcações em seqüência foi descartado, e 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.

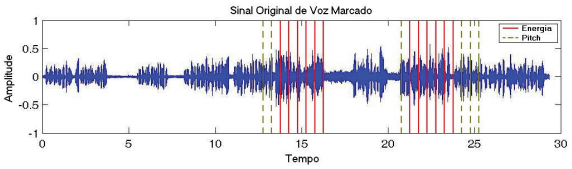

Figura 12 Sinal original de voz com as marcas de energia e de *pitch* que serão consideradas.

Num último estágio, o algoritmo de decisão une trechos de interesse que estejam muito próximos. Por exemplo, na Figura 12 há dois trechos separados por cerca de cinco segundos, o que pode indicar um único trecho de interesse pelo pequeno intervalo de tempo. Assim, realizando um estudo nos trechos marcados separadamente que fazem parte de um mesmo "bom momento" no Sinal I, descobriu-se que somente 5% desses intervalos foram maiores que oito segundos e que 80% foram menores que cinco segundos. Foi estipulado, então, um intervalo-limite de dez segundos a partir do qual trechos marcados separadamente são mantidos separados. Com este artifício aplicado ao Sinal I, os 53 trechos anteriormente marcados foram agrupados em apenas 24.

2 FUNCIONAMENTO DO SISTEMA

O método de classificação de eventos de interesse descrito na Seção 1 foi desenvolvido numa plataforma denominada MelhoresMomentos. O sistema foi desenvolvido em C++ com base em [8,9], utilizando MFC 8.0, biblioteca do Windows [10], IT++ 4.0.0, e a biblioteca para processamento de sinais [11], que utiliza a biblioteca MKL 9.1.027 da Intel [12].

A *interface* gráfica do sistema MelhoresMomentos é representada na Figura 13, cujas principais funcionalidades destacadas são:

(1) Sinal de vídeo sendo analisado;

(2) Barra de tempo deslizante para rápido avanço ou recuo do sinal de vídeo;

(3) Botões de "tocar" e "parar" o sinal de vídeo;

(4) Janela indicativa de marcação ou não do sinal sendo mostrado;

(5) Indicativo de início de trecho marcado;

(6) Indicativo de término de trecho marcado;

(7) Contador do trecho atual em relação ao total de segmentos;

(8) Lista de "bons momentos" detectados;

(9) Opção de limpeza da lista de trechos marcados.

Através do aplicativo, 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. O usuário ainda pode ajustar os limiares de energia e *pitch* para fazer um ajuste fino no desempenho do sistema.

Em termos de tempo de processamento, o sistema MelhoresMomentos necessitou de cerca de 3 minutos para detectar os melhores momentos de 45 minutos do Sinal I em um processador Intel Pentium Dual Core 3.06 GHz.

Figura 13 Interface gráfica do sistema MelhoresMomentos.

3 TESTES DE DESEMPENHO

Para uma avaliação mais criteriosa do método descrito na Seção 1, foram utilizadas duas medidas de desempenho: O percentual de "bons momentos" marcados corretamente (**%BMM**) e o percentual de trechos marcados que são efetivamente "bons momentos" (**%TMC**). A Figura 14 ilustra o que os parâmetros representam dentro dos resultados obtidos. A idéia é que o método marque todos os bons momentos, tendo um **%BMM** próximo de 100%, mesmo que alguns lances normais sejam também assinalados, gerando um **%TMC** abaixo de 100%.

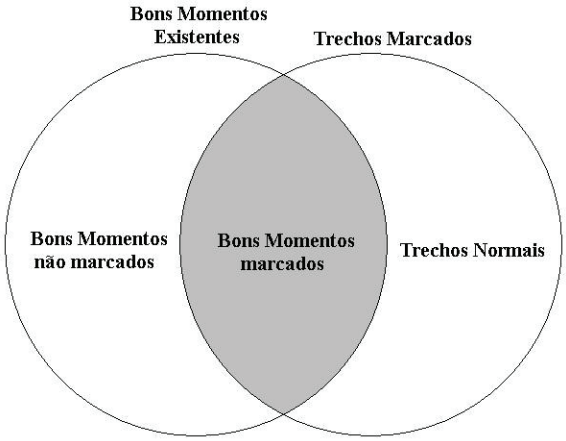

Figura 14 Diagrama de Venn ilustrando os parâmetros que servirão para visualização dos rsultados.

A Tabela 3 expressa os resultados iniciais para todos os sinais da base de dados descrita nas Tabelas 1 e 2. O Sinal I, que foi utilizado no desenvolvimento do método, teve todos os trechos desejados devidamente marcados, com alguns trechos indesejados também marcados. Na prática, para este sinal, cerca de 5 minutos de vídeo foram selecionados pelo método como contendo trechos de interesse. Os trechos indevidamente marcados podem ser eliminados de forma semi-automática por um operador humano. Esse processamento adicional, porém, fica extremamente facilitado pela alta seletividade do método, que reduziu o tempo de marcação em cerca de 90%. De modo geral, o mesmo desempenho se repetiu para todos os sinais do mesmo narrador atuando no Sinal I.

Tabela 3 Resultados iniciais do sistema MelhoresMomentos.

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	50,0

Nos sinais IV, V e VI, com locutores e captação diferentes do Sinal I, os resultados foram ruins, principalmente pelo fato de o **%BMM** ter sido abaixo de 100%. Isto se deve aos limiares utilizados para o Narrador I serem inadequados às caracteríticas de voz dos demais narradores. Realizando-se um ajuste empírico dos limiares de energia e *pitch*, de modo a se obter **%BMM** = 100%, obtêm-se os resultados indicados na Tabela 4.

Tabela 4 Resultados normalizados para diferentes narradores.

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

Mesmo com este ajuste, o funcionamento do sistema se manteve precário, já que os valores de **%TMC** tornaram-se extremamente baixos. Isto indica que um ajuste criterioso de todos os limiares determinados na Seção 1 deve ser feito para cada diferente narrador.

Procurou-se determinar a causa dos erros de classificação e percebeu-se que estes erros podem ser agrupados em três classes: (i) erros por emoção, onde o narrador aplica emoção à sua voz, porém em trechos descorrelacionados com a partida ou que não se caracterizam como um trecho de interesse, tais como anúncios, *replays* muito após o lance, início e término de partida etc.; (ii) erros devidos a outra pessoa, que são erros ocorridos em trechos de outros narradores, como comentaristas ou repórteres de campo; (iii) outros tipos de erros que não se encaixam nas duas categorias anteriores. A classificação dos erros ocorridos nos diferentes sinais é apresentada na Tabela 5.

Tabela 5 Distribuição dos erros por categorias.

Sinal	*Narrador*	*%Emoção*	*%Outra Pessoa*	*%Sem Motivo*
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

De modo geral, podemos concluir que o método aqui apresentado funciona muito bem como um detector de emoção do narrador para o qual o método foi treinado. Outros narradores, porém, requerem um ajuste dos limiares de classificação, para minimizar os erros pertencentes aos grupos (ii) e (iii).

Além da marcação correta dos "bons momentos", foi avaliado se o início e o fim dos bons momentos foram marcados satisfatoriamente (**%BMS**). Este tipo de análise

possui um caráter subjetivo, contando com a ajuda de um operador experiente. Os resultados indicados por este operador encontram-se na Tabela 6, que mostra que uma boa parcela dos trechos selecionados foi marcada satisfatoriamente. Na prática, percebeu-se que a principal razão de uma marcação inapropriada era a demora do narrador em aplicar emoção à voz. Aqui, mais uma vez, mostrou-se necessária a intervenção do usuário para redefinir os limites dos bons momentos que não foram marcados satisfatoriamente. Esta tarefa, porém, fica facilitada pelas funcionalidades presentes na *interface* gráfica da plataforma MelhoresMomentos.

Tabela 6 Percentual 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

4 CONCLUSÕES

Este artigo apresentou um método semi-automático de determinação dos melhores momentos de uma partida de futebol através do áudio do narrador. O método gera dois sinais de informação, um baseado na energia e outro no *pitch,* que realçam a possível ocorrência de "bons momentos". Um módulo de decisão utiliza ambas as informações para determinar os trechos de interesse, demarcando seus limites, e possivelmente agrupando trechos adjacentes correspondentes a um mesmo evento.

Os resultados foram satisfatórios, apesar de no estágio atual o método se mostrar dependente do locutor utilizado no seu desenvolvimento. A generalização do método exigiria um treinamento para cada narrador, montando-se um banco de narradores, ou ainda fazendo-se um ajuste automático dos limiares de decisão baseado em uma análise preliminar de curta duração.

Na opinião de um profissional de TV, com a generalização do método para outros narradores, será possível que um único operador seja responsável por editar os melhores momentos de diversas partidas que ocorram simultaneamente. De qualquer forma, em seu estado atual de desenvolvimento, o sistema MelhoresMomentos já é capaz de ser utilizado operacionalmente, de forma semi-automática, reduzindo o tempo de análise em cerca de 90% para os sinais com o mesmo narrador usado no seu desenvolvimento.

5 REFERÊNCIAS

[1] H. Christensen, Y. Gotoh, S. Renals, A Cascaded Broadcast News Highlighter , IEEE Trans. Audio, Speech, and Language Processing, 16(1), 1558-7916, Jan. 2008.

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

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

[4] Wikipedia, [http://en.wikipedia.org/wiki/Circular_buffer]

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

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

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

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

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

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

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

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