Detecção de Refrão em Sinais de Áudio usando Extração de Características de Intensidade do Som

Renato Celso Santos Rodrigues, Geber Ramalho, Giordano Cabral

Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Caixa Postal 7851 – 50.732-970 – Recife – PE – Brasil

{rcsr, glr, grec}@cin.ufpe.br

Abstract. *This paper proposes a paradigm shift for chorus detection, based on the exploitation of the time domain rather than the frequency domain. State of the Art methods generally segment the signal by exploiting the presence of musical notes to measure the similarity between the music sections with Euclidean distance between Chroma and MFCC vectors. The proposed method eliminates the segmentation step and extracts the envelope of the signal, measuring the similarity between the envelopes of the parts of the song by correlation function, rendering the method independent of the presence of musical notes in the signal. It was carried out a comparative study of current approaches and the proposed method, highlighting their differences, advantages and disadvantages. The results indicate that the use of correlation function for the envelope signal achieves hit rate and performance at the same magnitude of current methods.*

Resumo. *Este trabalho propõe uma mudança de paradigma para a solução do problema da detecção de refrão, baseada na exploração do domínio do tempo em lugar do domínio da frequência. Métodos do Estado da Arte geralmente segmentam o sinal, explorando a presença de notas musicais para medir a similaridade entre os trechos da música com distância euclidiana entre vetores Chroma e MFCC. O método proposto elimina a etapa de segmentação e extrai a envoltória do sinal, medindo a similaridade entre as envoltórias das partes da música por função de correlação, tornando o método independente da presença de notas musicais no sinal. Foi realizado estudo comparativo entre as abordagens atuais e o método proposto, destacando suas diferenças, vantagens e desvantagens. Os resultados indicam que a utilização de função de correlação sobre a envoltória do sinal indica obtém taxa de acertos e desempenho na mesma ordem de grandeza dos métodos atuais.*

1. Introdução

Segundo Goto (2006), o refrão (*chorus* ou *refrain*, na literatura) é a parte mais repetida e memorável de uma música. É o trecho mais proeminente e representativo de uma canção, mais facilmente reconhecível e memorizável pelos ouvintes. Desta forma, o refrão é capaz de representar a música em um contexto de rápida visualização de álbuns musicais, como em aplicações de navegação ou recuperação de música.

Uma função de Preview oferece ao ouvinte uma pré-visualização de um álbum com pequenos trechos de suas músicas. É necessário oferecer ao ouvinte o trecho mais representativo de cada música, sendo muito útil para isso a detecção de refrão. Neste contexto, o objetivo de um sistema de detecção de refrão é identificar os pontos

extremos do refrão no eixo temporal ao longo da música: o instante no qual ele começa; e o instante no qual ele termina.

Entretanto, detecção de refrão não é simples. Embora seja intuitivo ao ser humano identificar o trecho mais representativo de uma música, não é fácil definir precisamente onde ele começa e onde termina. Para um computador, esta imprecisão na marcação dos extremos do refrão pode causar uma diminuição da taxa de acertos, sobretudo quando se utiliza como métrica de sucesso uma medida baseada na comparação entre o refrão retornado pela aplicação com o refrão anotado, previamente definido em uma marcação manual.

Outra dificuldade é que a detecção de refrão é baseada na identificação de partes similares entre si ao longo da música, mas uma ocorrência do refrão pode apresentar diferenças em relação às demais. Dependendo da forma como esta similaridade é medida, as diferenças entre as ocorrências dos refrãos, como a presença ou ausência de instrumentos específicos, ou a inclusão de arranjos musicais distintos, ou ainda diferenças na intensidade do som, podem reduzir o grau de similaridade entre um par de ocorrências do refrão na música.

Uma solução para o problema da detecção de refrão deve ter baixa taxa de erros e alto desempenho. Deve ser também robusta, com uma taxa de acertos e desempenho que não variem em função de características peculiares ao estilo musical, como o ritmo, por exemplo.

Neste contexto, o objetivo deste trabalho é propor uma abordagem alternativa para a detecção de refrão, testando-a experimentalmente e realizando um estudo comparativo, destacando suas diferenças, vantagens e desvantagens em relação às soluções do Estado da Arte. Esta mudança de paradigma é realizada pela substituição de alguns dos passos dos métodos de Goto e Eronen por outros que explorem formas diferentes de executar a mesma tarefa.

2. Estado da Arte

As duas principais abordagens para solução do problema da detecção de refrão são os trabalhos de Goto (2006) e Eronen (2007). Ambas começam com a segmentação do sinal, dividindo-o em fragmentos que serão comparados entre si, medindo-se o grau de similaridade entre cada par destes segmentos. Este processo realiza uma busca pelas batidas da música, mais evidentes em função dos instrumentos percussivos, detectadas conforme trabalho de Seppänen (2006), onde cada segmento da música é o trecho entre duas batidas consecutivas. Esta técnica obtém o menor número possível de fragmentos do sinal que ainda preserva a precisão de uma solução baseada na extração da frequência dos sons, reduzindo o seu custo nas etapas seguintes. Porém, o desempenho da solução torna-se uma função da frequência das batidas da música, porque músicas mais aceleradas serão divididas em um número maior de segmentos.

Realiza-se então a extração de características musicais de cada segmento, úteis no cálculo da similaridade entre cada dois segmentos do sinal. As características adotadas foram o Chroma Vector para ambos os métodos e ainda o vetor MFCC para o método de Eronen, ambos baseados na frequência do sinal em cada segmento. O Chroma Vector, construído conforme o trabalho de Goto (2006), possui 12 posições, uma para cada Pitch Class, que contém a soma das magnitudes das frequências de cada

Pitch Class em 6 oitavas, no espectro obtido por FFT em uma janela deslizante ao longo do segmento.

Depois, usa-se distância euclidiana para medir as similaridades entre estes vetores, onde uma menor distância significa uma maior similaridade. Uma desvantagem neste passo é a necessidade de tratar as modulações (mudanças na tonalidade da música em tempo de execução), porque elas anulam a similaridade entre refrãos executados em tonalidades diferentes. Este tratamento acrescenta um pouco de complexidade e custo ao algoritmo da solução, embora não seja difícil de ser realizado.

Após medir as distâncias, Eronen (2007) armazena-as em uma matriz de similaridade, que terá em cada posição Aij da matriz o valor da similaridade entre os segmentos i e j do sinal. Goto (2006) utiliza um triângulo similar, que armazena a similaridade entre dois trechos em função também da distância entre eles no tempo da música. Uma diagonal de alta similaridade nesta matriz (ou uma linha de alta similaridade no triângulo) identifica uma repetição entre dois trechos da música. Esta técnica permite a utilização de filtros de processamento de imagens para refinar a detecção destas diagonais (ou linhas) pelo realce de seus valores em relação aos vizinhos, ou pela eliminação de valores espúrios. A Figura 1 traz um exemplo de matriz de similaridade.

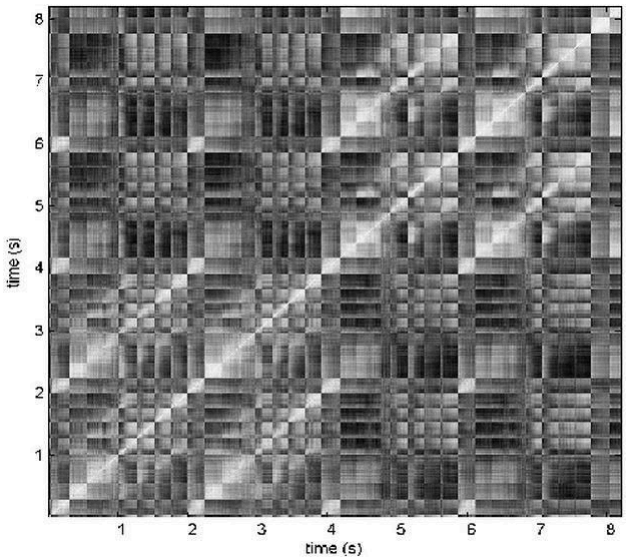

Figura 1. Exemplo de matriz de similaridade.

No final de ambos os métodos são identificadas as repetições através da dicotomização da matriz (ou triângulo) com a anulação de valores abaixo de um limiar. Depois, sobre cada repetição são aplicadas heurísticas que calculam a probabilidade dela ser uma correspondência entre dois refrãos da música. A repetição com maior probabilidade terá um de seus trechos apontado como o refrão da música.

As heurísticas variam de abordagem para abordagem, e dependem de observações realizadas pelo autor do método sobre os refrãos das canções. Eronen (2007) adotou quatro heurísticas: uma repetição que contém refrãos provavelmente encontra-se próximo de um quarto e três quartos do tempo da música, possui alta similaridade média, alta intensidade média, e um dos trechos presente em outras repetições. Goto (2006) adotou três heurísticas: uma repetição que contém refrãos

provavelmente tem entre 8 e 40 segundos de duração, é inclusa em outra repetição maior e contém duas outras repetições menores.

Ambas as abordagens operam com uma taxa de acerto entre 80% e 85% em bases de dados específicas a cada autor. A base usada por Goto é constituída por 100 músicas populares, enquanto a usada por Eronen tem 206 músicas populares e rock. Os métodos de ambos trabalham retornando o refrão em 10 a 15 segundos para uma música de aproximadamente 3 a 4 minutos, para um computador de configuração mediana, acessível a um usuário comum.

3. O Método

A principal mudança proposta no método em relação às abordagens de Goto e Eronen é a substituição da extração de características do Chroma Vector e do vetor MFCC pela extração da envoltória do sinal, que é a curva que descreve a variação de energia do sinal ao longo do tempo. Esta mudança torna o método independente das frequências das notas musicais do sinal, dispensando o tratamento de modulações.

O sinal de áudio musical tem sua envoltória modulada por ondas na faixa de frequência audível pelo ouvido humano, resultante das frequências de suas notas. Cada vez que um trecho da música é executado, não apenas as frequências das notas musicais estarão presentes, mas também a contribuição de cada nota musical para a variação de amplitude do sinal. Assim, a envoltória do sinal também conserva a similaridade existente entre os trechos que se repetem, independente da frequência que a modulou.

Para medir a similaridade entre os trechos da música, o método utiliza o máximo da função de Correlação (Equação 1), uma vez que a filtragem das frequências de modulação do sinal impossibilita a utilização da distância euclidiana entre vetores Chroma e MFCC. Esta função tem máximos quando, para um dado lag τ, o sinal x é similar ao sinal y.

$$R_X(\tau) = \int_{t'}^{t'+T} x(t+\tau)\, y(t)dt \text{ , t' constante} \qquad (1)$$

Outra mudança proposta é a eliminação da etapa de segmentação, que simplifica o método, remove o custo empregado na detecção de batidas, torna o seu tempo de resposta independente da frequência destas batidas, e possibilita uma maior precisão no corte dos extremos do refrão, que não estará mais limitado às batidas do sinal.

O restante desta seção descreve a sequência de passos do método proposto.

3.1. Extração da Envoltória do Sinal de Áudio

A extração de Envoltória do sinal utiliza uma janela deslizante que se desloca ao longo do sinal recuperando o valor máximo deste frame em cada iteração. A sequência destes valores é armazenada em outro sinal que, com o devido ajuste do tamanho da janela e do valor do deslocamento, conterá a envoltória do sinal de áudio, que é a descrição aproximada da variação de sua energia ao longo do tempo, mas com um número de amostras menor. O tamanho do sinal resultante será aproximadamente o valor da razão entre o tamanho do sinal de áudio pelo deslocamento da janela adotado, ambos medidos em números de amostras digitais do sinal original. A Figura 2 possui um gráfico com 8 notas consecutivas tocadas por um violão no domínio do tempo. O sinal de envoltória obtido sobre este sinal encontra-se no gráfico da Figura 3.

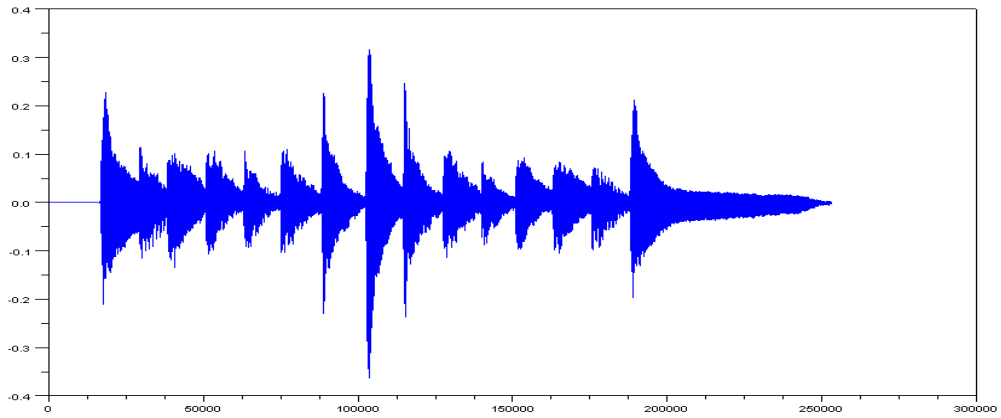

Figura 2. Sinal de áudio com 15 notas de um violão.

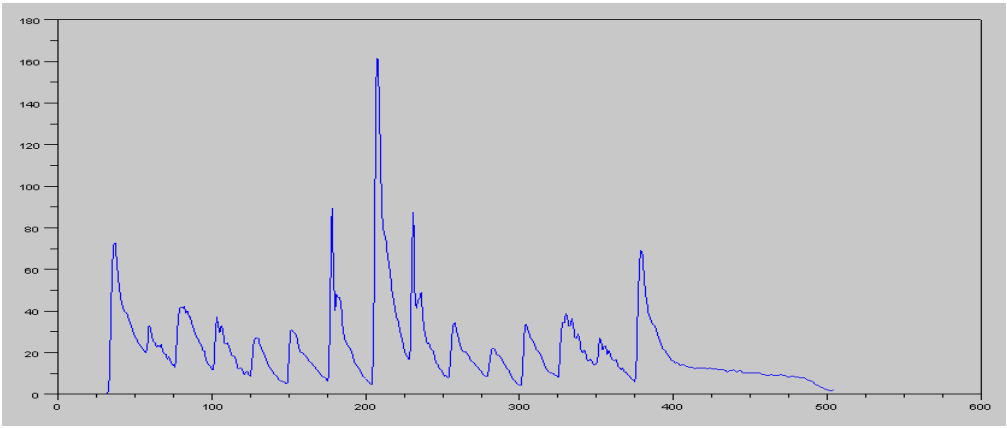

Figura 3. Envoltória do sinal de áudio do Gráfico 2.

Observando o eixo horizontal de ambos os gráficos nota-se uma redução no tamanho do sinal de aproximadamente três ordens de grandeza, o que permite uma significativa redução de custos nas etapas seguintes deste método.

Quanto maior for o valor adotado para o deslocamento da janela deslizante, menor será o sinal de envoltória obtido, e quanto menor for o sinal de envoltória obtido, menor será o custo do processamento realizado sobre ele nas etapas seguintes do método. Entretanto, um deslocamento muito grande pode resultar em perda de informação da similaridade presente na envoltória, o que é indesejável. O valor do deslocamento adotado na abordagem foi 50 milisegundos, obtido empiricamente observando-se um bom compromisso entre a redução do tamanho e a preservação da similaridade.

Reduzir o tamanho da janela deslizante reduz o custo da construção do sinal de envoltória, mas aumenta a frequência de corte na obtenção da envoltória**,** permitindo que frequências mais baixas de notas musicais (mais graves) insiram ruídos na envoltória obtida. Mas uma janela muito grande perde em precisão, podendo incluir muitos padrões de variação de amplitude (ataque, decaimento, sustentação e relaxamento) consecutivos em uma só posição do sinal resultante, perdendo informação. Diante deste problema, o valor adotado foi 10 milisegundos.

3.2. Construção da Matriz de Similaridade

A matriz de similaridade é construída como no método de Eronen (2007), mas o cálculo de seus valores é diferente. Em cada posição da matriz de similaridade é armazenado, como métrica de similaridade, o máximo do sinal de correlação obtido tendo como entrada duas janelas deslizantes, que se deslocam ao longo do sinal de envoltória. O tamanho da janela e do deslocamento adotados foram 10 segundos e 1 segundo respectivamente (medidos em amostras do sinal de envoltória), obtidos empiricamente preservando um bom compromisso entre precisão, custo e taxa de acertos.

Para reduzir o custo do método nesta etapa é preenchido apenas do triângulo superior da matriz de similaridade. Uma vez que ela é simétrica, apenas um dos triângulos é necessário e suficiente para os processamentos posteriores, que somente utilizarão esta metade da matriz. A matriz à esquerda da Figura 3 é resultante deste processo sobre um sinal de áudio musical de aproximadamente 4 minutos.

3.3. Extração das Diagonais de Similaridade

As técnicas utilizadas neste passo são praticamente as mesmas utilizadas na abordagem de Eronen. A primeira é a filtragem de nitidez, que utiliza um kernel quadrado de 25 posições que se desloca por todo o triângulo superior da matriz e aumenta o valor da posição central do kernel em relação às demais se o seu valor de similaridade máxima estiver na sua diagonal principal, e diminui o valor deste elemento central em relação aos demais em caso contrário. Ao fim deste processo, espera-se que as diagonais de mais alta similaridade estejam mais nítidas em relação aos demais valores da matriz. Na Figura 4, a matriz de similaridade à direita é resultante desta filtragem de nitidez sobre a matriz à esquerda.

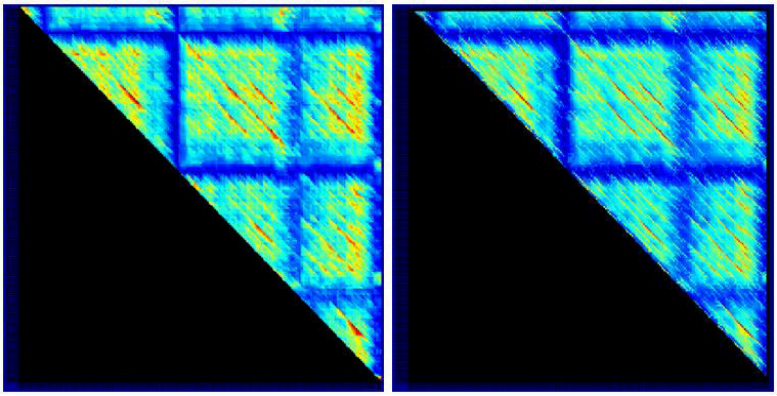

Figura 4. Matrizes de similaridade antes (E) e depois (D) da filtragem de nitidez.

A seguir calcula-se a similaridade média de cada diagonal do triângulo superior da matriz, e estes valores são armazenados. Em seguida, as 10 diagonais de maior média são selecionadas, e as demais são cortadas da matriz pela anulação de todos os seus valores. Após este corte, é calculado um limiar que permita a dicotomia de todos os valores não nulos onde exatamente 20% destes valores permaneçam acima deste limiar e o restante seja anulado. Isto é obtido pela concatenação das 10 diagonais não nulas em uma só sequência de valores, seguida pela ordenação decrescente destes valores (sort), escolhendo-se como limiar o valor correspondente à posição que tem como índice o tamanho da sequência dividido por cinco. Obtido este limiar, todo valor abaixo dele é também anulado. Ao fim deste processo espera-se que somente os trechos de mais alta

similaridade nas diagonais de maior similaridade média estejam não nulos na matriz. A escolha de 20% para este limiar de corte foi empírica, e realizada no trabalho de Eronen. A matriz da esquerda da Figura 5 é resultante deste processo de corte quando aplicado sobre a matriz à direita na Figura 4.

A última filtragem sobre a Matriz de similaridade é a remoção de gaps nulos entre diagonais não nulas, provenientes de curtas diferenças existentes entre os refrãos. São removidos os gaps inseridos em um intervalo não nulo de pelo menos 12 segundos, e que não ultrapassem o limite máximo de 4 segundos. Estes valores foram obtidos a partir dos valores sugeridos no trabalho de Eronen, originalmente medidos em batidas da música, sendo ajustados empiricamente em testes.

No final do processo descrito nesta seção espera-se ter as diagonais de repetições bem definidas na matriz de similaridade, cada uma contendo a correspondência entre dois trechos de música similares entre si.

3.4. Seleção da repetição candidata a refrão

Neste passo, as repetições encontradas na matriz de similaridade na etapa anterior são listadas pelo armazenamento de seus extremos (instante inicial e instante final). O número típico de repetições listadas está entre 10 e 15. Dentre estas repetições, que podem incluir tanto um refrão como outra parte da música que também é repetida, somente uma deve ser retornada pelo método como refrão obtido. Tal como nas soluções Goto e Eronen, esta seleção é realizada por heurísticas.

A primeira heurística adotada nesta solução é a do range de tamanho de refrãos. O range adotado com base em observações exaustivas sobre músicas de diversos estilos musicais é o intervalo de 5% a 25% da música, onde repetições fora deste range são descartadas, pois provavelmente possuem muitos trechos que não fazem parte do refrão. Na Figura 5, a matriz à direita é resultante deste processo sobre a matriz à esquerda, juntamente com o processo de remoção de gaps nulos descrito na seção 3.3. Esta heurística é a única que elimina uma repetição dentre as candidatas a refrão.

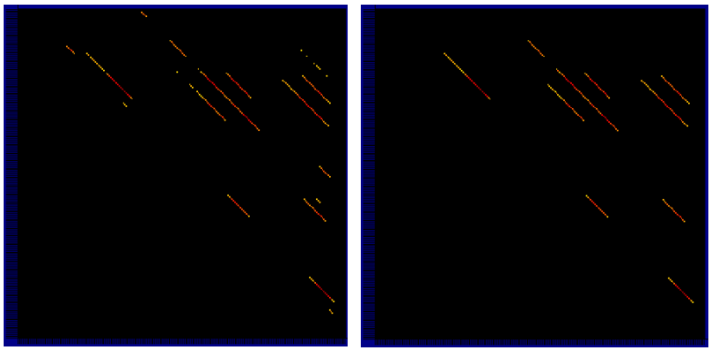

Figura 5. (E) Matriz após corte de 80% dos valores; (D) Matriz após remoção de gaps nulos e repetições fora do range.

As heurísticas descritas a seguir foram implementadas na forma de funções que retornam um valor normalizado (entre zero e um) interpretado como a probabilidade da repetição avaliada conter um refrão.

A segunda heurística testada foi o somatório do grau de correspondência com as outras repetições listadas, onde a repetição que tiver maior soma do grau de correspondência com todas as demais repetições é uma forte candidata a refrão. A

correspondência entre duas repetições é calculada usando F-measure entre os seus respectivos segmentos, tanto no eixo horizontal como no eixo vertical da matriz. A F-Measure, cujo cálculo é descrito na seção 4, retorna um valor entre zero (interseção nula) e um (intervalos idênticos) para os dois segmentos de entrada, identificados pelos seus extremos no eixo do tempo.

Uma matriz de correspondência é criada, equivalente à matriz de similaridade, e cada posição armazenará a correspondência de cada par de diagonais entre si, onde cada triângulo da matriz armazena as F-measures em relação a um dos eixos. Assim, a medida desta heurística, para cada repetição, é a soma dos valores contidos na linha com os valores contidos na coluna desta posição, relativas às correspondências tanto no eixo horizontal como no eixo vertical. Este valor, no fim, é normalizado, para que seja retornado um valor entre zero a um. Esta heurística considera o número de repetições de um trecho do sinal ao longo da música, porque um segmento com maior número de repetições tem maior chance de ser um refrão, pela própria definição de refrão.

A terceira heurística é a Similaridade Média, onde a diagonal cuja similaridade média na matriz de similaridade for maior tem maior probabilidade de ser o refrão. Desta forma, as similaridades médias de cada diagonal são calculadas e depois os valores são também normalizados.

A quarta heurística considerada é a do último segmento. Observando diversas músicas de diversos estilos musicais, facilmente se nota que é comum a música terminar com uma sequência de ao menos dois ou três refrãos consecutivos. Desta forma, uma repetição que contém um segmento localizado na região final da música tem grandes chances de conter um refrão. Neste caso, a probabilidade do segmento conter um refrão é inversamente proporcional à distância de seu ponto médio ao final da música.

A quinta heurística considerada é a da localização da repetição, onde a repetição que seja combinação do trecho mais próximo a um quarto combinado com o trecho mais próximo a três quartos do sinal de música mais provavelmente tem um refrão. A probabilidade aqui é caculada de maneira similar à quarta heurística, mas a distância considerada aqui é a do ponto médio do segmento até o ponto equivalente a ¼ e ¾ da música no eixo do tempo.

Nos testes realizados, combinações destas heurísticas foram também testadas. A seção a seguir descreve o procedimento de testes, bem como os resultados alcançados e uma discussão sobre estes resultados.

4. Experimento

A base de dados utilizada possui 50 canções da música gospel brasileira, que recebe influências como a da música americana e de diversos ritmos nacionais e regionais, incorporando estilos como o soul, o rock em diversas vertentes, o pop, o romântico, o samba, o forró, conferindo uma variedade razoável. As músicas foram ouvidas e os refrãos foram marcados manualmente associando-se a cada canção um identificador e os pares de extremos (instante inicial e instante final) de cada refrão encontrado. As músicas comtemplam ampla variedade de instrumentos musicais, tendo também solos em voz masculina, feminina, músicas com harmonia de vozes, ou ainda cantadas juntamente com as pessoas da plateia.

A métrica de sucesso é a mesma do Estado da Arte, a F-Measure, definida como a média harmônica entre a taxa de Recall (R) e Precisão (P). Seja *a* o refrão anotado manualmente e *b* o refrão retornado pela solução. Primeiro se mede a interseção entre estes dois intervalos *Intersect(a,b)*, e depois é calculado o valor de R, razão entre *Intersect(a,b)* e o tamanho de *a* (Equação 2), e o valor de P, razão entre *Intersect(a,b)* e o tamanho de *b* (Equação 3). Depois se calcula a média harmônica de R e P, razão entre o produto de R e P e a soma de R e P (Equação 4), que é a taxa de acerto.

$$R = \frac{Inter\,sect(a,b)}{length(a)} \quad (2)$$

$$P = \frac{Inter\,sect(a,b)}{length(b)} \quad (3)$$

$$F\_measure = \frac{2RP}{R+P} \quad (4)$$

A F-Measure retorna um valor entre zero e um que representa a correspondência no tempo entre *a* e *b*, onde zero significa correspondência nula (interseção inexistente) e um indica correspondência máxima, caso onde *a* possui exatamente os mesmos extremos (início e fim) que *b*.

Para cada música da base, o F-Measure foi calculado, para cada heurística ou combinação de heurísticas que foi testada, e os resultados obtidos foram registrados para análise posterior. Ao fim do processamento é calculada a F-Measure média para cada heurística ou combinação de heurísticas, e esta será a taxa de acertos final. As taxas de acerto obtidas foram comparadas com o caso ótimo, que assume que as heurísticas sempre acertam o segmento ótimo que maximiza o valor da F-Measure para uma música, e com o pior caso, que é a escolha aleatória de um segmento dentre os listados imediatamente antes da etapa de seleção de refrão.

Considerando as heurísticas descritas na seção 4.4, que são "Grau de Correspondência", "Similaridade Média", "Último Segmento", "Localização da Repetição", representadas respectivamente por h1, h2, h3 e h4, os testes apresentaram os resultados para o F-Measure descritos na Tabela 1.

Tabela 1. Resultados obtidos com F-Measure média.

Heurística	F-Measure Média	Heurísticas	F-Measure Média
Caso Ótimo	89,92%	h1 e h2	82,49%
Randômico	67,35%	h1 e h3	79,29%
h1	82,09%	h1 e h4	81,89%
h2	82,09%	h2 e h3	71,15%
h3	68,21%	h2 e h4	71,15%
h4	68,54%	h3 e h4	71,15%

Durante os testes de custo computacional, uma música de 4 minutos e 3 segundos foi escolhida para testar o tempo de resposta da aplicação, obtendo-se o tempo de resposta de 13 segundos.

4.1. Avaliação

Algumas heurísticas ou combinação de heurísticas atingiram uma taxa de acertos entre 80% a 85%. O custo computacional do método proposto é função apenas do tamanho da

entrada, pois a razão entre o tamanho da música e o tempo de resposta é sempre mantida. Esta regra não é observada no Estado da Arte, onde o desempenho é função do número de batidas da música e, portanto, de sua velocidade. Como o desempenho obtido no Estado da Arte foi de cerca de 10 segundos para músicas de 3 a 4 minutos, foi comprovado que o custo computacional é da mesma ordem de grandeza dos métodos atuais, considerando um computador com uma configuração média.

A heurística que obteve resultado mais próximo do ótimo nos experimentos foi a h1 (Grau de Correspondência). A Figura 6 apresenta a mesma matriz à direita da Figura 5 com a repetição que foi selecionada por esta heurística destacada em azul, onde esta heurística obteve para esta música uma F-Measure de 92,3%.

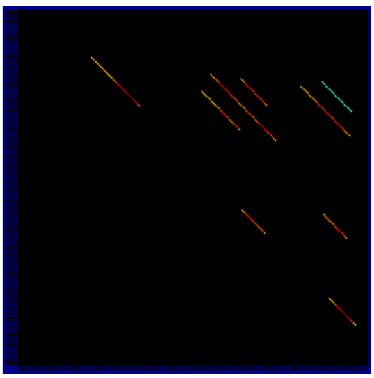

Figura 6. Matriz com repetição selecionada destacada.

5. Conclusão

Os resultados comprovaram que a utilização da função de correlação sobre o sinal de envoltória do sinal para estimativa da similaridade entre trechos da música também é eficaz para a detecção de refrão.

A abordagem deste trabalho realiza a detecção de refrão de uma forma diferente, com suas vantagens e desvantagens específicas. Uma vantagem é a independência da presença de notas musicais, dispensando a exploração do domínio da frequência, sujeita a dificuldades como a modulação, que exigem uma lógica de tratamento adicional. Além disso, espera-se que soluções dependentes de notas musicais não funcionem bem na estimativa de partes que se repetem ao longo do tempo em sinais de áudio que não tenham notas musicais, o que em tese não seria problema para o método proposto.

Explorar o domínio da frequência não é necessariamente pior ou mais complexo que explorar a intensidade do som em uma solução de detecção de refrão. Abordagens diferentes possuem vantagens e desvantagens diferentes que poderão ser mais adequadas a cada tipo particular de aplicação.

Trabalhos futuros incluem encontrar dinamicamente o limiar de dicotomização da matriz de similaridade, atualmente fixo em 20%, e a exploração mais profunda das heurísticas empregadas na etapa de seleção de refrão, refinando as que já existem ou encontrando novas formas de se eleger uma repetição candidata.

Referências

Bartsch, M. A. and Wakefield, G. H. (2005) "Audio Thumbnailing of Popular Music Using Chroma-Based Representation", In: IEEE Trans. on Multimedia, vol. 7, no. 1, Feb. 2005, pp. 96-104.

Cooper, M. and Foote, J. (2003) "Summarizing Popular Music Via Structural Similarity Analysis," In: Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2003, October 19-22, 2003, New Paltz, NY.

Ellis, D. (2006) "Beat Tracking with Dynamic Programming", MIREX 2006 Audio Beat Tracking Contest system description, Sep 2006, available athttp://www.ee.columbia.edu/~dpwe/pubs/Ellis06-beattrack.pdf

Eronen, A. (2007) "Chorus Detection with Combined use of MFCC and Chroma Features and Image Processing Filters", in Proc. of the 10th International Conference on Digital Audio Effects (DAFx-07),Bordeaux, France, September 10-15, 2007.

Goto, M. (2006) "A Chorus Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station", In: IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 5, Sept. 2006 pp. 1783 – 1794.

Levy, M., Sandler, M. and Casey, M. (2006) "Extraction of High-Level Musical Structure From Audio Data and Its Application to Thumbnail Generation", In: Proc. IEEE ICASSP 2006, vol. V, pp. 13-16.

Marolt, M. (2006) "A Mid-level Melody-based Representation for Calculating Audio Similarity", In: Proc. of the 7th International Conference on Music Information Retrieval, ISMIR 2006, Victoria, Canada, 8 - 12 October 2006.

Paulus, J. and Klapuri, A. (2006) "Music Structure Analysis by Finding Repeated Parts", In: Proc. of the 1st Audio and Music Computing for Multimedia Workshop (AMCMM2006), Santa Barbara, California, USA, October 27, 2006, pp. 59-68.

Seppänen, J., Eronen, A. and Hiipakka, J. (2006) "Joint Beat & Tatum Tracking from Music Signals", In: Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR), Victoria (BC), Canada, October 8-12, (2006).

Shiu, Y., Jeong, H. and Kuo, C. (2006) "Similarity Matrix Processing for Music Structure Analysis", In: Proc. of the 1st Audio and Music Computing for Multimedia Workshop (AMCMM2006), October 27, 2006, Santa Barbara, California, USA.