Sociedade de Engenhariade Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, São Paulo, SP

Transcrição automática de sinais de áudio monofônico baseada em quadro do Artigo

Tiago Fernandes Tavares[1], Jayme Garcia Arnal Barbedo[1,2] e Amauri Lopes[1]
[1]Departamento de Comunicações – FEEC – UNICAMP
Caixa postal 6101 – CEP 13.083-970 – Campinas – SP
[2]Harvard School of Applied Sciences – Harvard University
ZIP 02138 – Cambridge – MA – USA
ttavares@fee.unicamp.br, {jgab, amauri}@decom.fee.unicamp.br

RESUMO
Este artigo apresenta um novo algoritmo para transcrição automática de sinais de áudio, o qual utiliza quadros de tamanho variável de forma a otimizar o compromisso entre a resolução temporal do processo de transcrição e a menor freqüência identificável pelo algoritmo. O processo de determinação do tamanho do quadro se apóia em características específicas do método Yin, embora possa ser modificado para operar em conjunto com outros métodos de extração da freqüência fundamental. Os resultados obtidos demonstram que a técnica pode ser aplicada com sucesso na transcrição de sinais monofônicos. A extensão do método a sinais polifônicos será abordada em estudos futuros.

0 INTRODUÇÃO

Processos de transcrição automática de música têm por objetivo a construção de alguma forma de notação musical a partir de um sinal sonoro digitalizado empregando o processamento de sinais por computador. Quando lidamos com áudio monofônico, a determinação da nota musical tocada num determinado intervalo de tempo envolve a determinação da freqüência fundamental da onda sonora nesse intervalo [1,2]. Por esse motivo, o processo de transcrição automática de áudio normalmente começa com a divisão do sinal analisado em quadros de tamanho conhecido, seguida pelo cálculo da freqüência fundamental em cada um deles. Sabendo os tempos de início e término de cada freqüência fundamental, pode-se determinar durações de notas e, assim, obter uma notação musical simplificada (estilo MIDI) correspondente à execução analisada [1,2].

Os algoritmos mais comuns para detecção de freqüência fundamental precisam, para seu funcionamento, que o quadro tenha pelo menos duas vezes a duração do maior período analisado [3,4,5]. O aumento na duração do quadro, porém, significa que um trecho mais longo de áudio será

considerado uniforme quanto à freqüência fundamental, fazendo com que o limite inferior de duração de notas detectáveis seja também modificado. Desta forma, a divisão do sinal de áudio em quadros implica em um compromisso entre a menor freqüência e a menor duração de uma nota detectável. Neste artigo, é proposto um algoritmo para variar o tamanho do quadro buscando otimizar o compromisso citado.

Foi estudado o comportamento de um método de detecção de freqüência fundamental já existente quando aplicado sobre um quadro de duração menor que o dobro do período fundamental do sinal analisado. Conhecendo esse comportamento, pode-se determinar quando o quadro de análise tem duração muito curta para que o método de detecção de freqüência fundamental atue corretamente sobre o trecho avaliado. Ao detectar a ocorrência desse comportamento, o algoritmo decide que a duração do quadro deve ser aumentada. Em contrapartida, sempre que o período fundamental detectado for menor que a metade da duração do quadro, este pode ser reduzido para que se ajuste à onda analisada.

1 DETECÇÃO DA FREQÜÊNCIA FUNDAMENTAL

O primeiro passo para a construção de um transcritor automático de áudio monofônico é a escolha de um método de detecção de freqüência fundamental. Existem vários métodos, entre eles o *Harmonic Product Spectrum* (HPS)[4], a detecção baseada em autocorrelação[1,3,4], análise de coeficientes cepstrais[3] e o método Yin[3,5], que, por apresentar o melhor desempenho[5], foi escolhido como base para para este trabalho.

O método Yin, cujo nome deriva de *Yin-Yang*, determina a freqüência fundamental a partir da função diferença, definida pela Equação 1.

$$d_t[n] = \sum_{k=1}^{N} (x[k] - x[k+n])^2 \qquad (1)$$

A função diferença mede o quanto um sinal é diferente dele mesmo deslocado no tempo. Sinais periódicos, quando deslocados de tempo igual ao seu período fundamental, se assemelham muito ao sinal original, fazendo com que a função diferença apresente mínimos locais em pontos correspondentes a períodos fundamentais da onda analisada.

Como a função pode apresentar mais de um mínimo local, sendo apenas o primeiro deles o desejado – excluindo-se o mínimo referente ao deslocamento zero – o método Yin propõe que seja calculada a *função diferença absoluta acumulada normalizada (FDAAN)*, definida pela Equação 2.

$$d'_t[n] = 1 \; para \; n = 0 \; ;$$
$$d'_t[n] = \frac{d_t[n]}{\frac{1}{n}\sum_{j=1}^{n} d_t[j]} \; para \; n \neq 0 \qquad (2)$$

A primeira expressão da Equação 2 elimina o primeiro vale da função diferença. Para encontrar o mínimo, então, procura-se o primeiro valor de *n* para o qual a *FDAAN* é um mínimo local menor que um certo limite absoluto (definido experimentalmente). É possível, porém, que nenhum ponto satisfaça as condições de parada. Nesse caso, a busca retorna o índice do menor valor da *FDAAN*.

Em qualquer circunstância, o ponto que foi determinado como mínimo, junto a seu antecessor e seu sucessor, são interpolados por uma parábola. Adota-se como mínimo da função o valor mínimo dessa parábola, com o objetivo de aumentar a resolução da detecção de freqüência.

2 A DURAÇÃO DO QUADRO DE ANÁLISE

A transcrição baseada no método Yin baseia-se no fato de que, para pequenos trechos, um sinal de áudio monofônico gerado por um instrumento harmônico é periódico. Ao se dividir o sinal em quadros com duração suficientemente pequena, pode-se extrair a freqüência fundamental em cada um deles. Esse procedimento permite que seja inferido o início e o final de cada nota musical executada, permitindo que seja obtida a notação musical correspondente.

Para determinar a duração do quadro, alguns aspectos devem ser levados em consideração. Quando é utilizado o método Yin, a menor freqüência fundamental detectável é aquela correspondente a um período igual à metade da duração do quadro, de forma que um quadro muito pequeno levará à perda de notas graves. Por outro lado, o aumento na duração do quadro faz com que um trecho maior de áudio seja analisado a cada iteração, reduzindo a resolução temporal do processo de transcrição e aumentando o esforço computacional demandado pelo algoritmo de detecção de freqüência fundamental.

Visando otimizar o compromisso entre a resolução no tempo e o alcance em freqüência, foi desenvolvido um algoritmo capaz de variar a duração do quadro de acordo com a demanda de cada trecho do sinal analisado. Esse algoritmo se apóia em características específicas do método Yin, como será visto adiante, embora possa ser redesenhado para outros métodos de extração de freqüência fundamental.

A redução da duração do quadro pode acontecer sempre que o período fundamental detectado tiver duração menor que a metade da duração do quadro, mantendo a condição de funcionamento do algoritmo. O aumento da duração do quadro depende de fatores que devem ser analisados mais cautelosamente. Quando é aplicado o método Yin, a busca por um mínimo local na *FDAAN* pode não encontrar valor abaixo do limite estabelecido, o que significa que a onda analisada não apresenta grande semelhança consigo mesma quando deslocada no tempo, caso em que é retornado o índice do mínimo global da função. Essa inexistência de semelhança pode significar que a duração do quadro não é suficiente para conter dois períodos fundamentais da onda analisada. A partir dessa característica, pode-se determinar que quando o método Yin retorna um mínimo global acima do limiar adotado o quadro deve ter sua duração aumentada.

A quantidade de aumento ou diminuição da duração quadro é, em princípio, arbitrária, embora quantidades muito grandes possam levar o algoritmo a oscilações na duração do quadro e quantidades muito pequenas possam fazer a variação aplicada tornar-se cada vez mais imperceptível. Neste trabalho, a duração é multiplicada ou dividida por dois, de forma a variar o limite inferior de freqüência detectável em uma oitava sempre que isso se fizer necessário.

O processo de transcrição deve ser iniciado com a escolha de uma duração inicial para o quadro. Uma vez que essa duração convergirá para um valor considerado ideal pelo algoritmo de variação para cada nota musical detectada, essa

duração pode ser escolhida livremente dentro de limites razoáveis, da mesma ordem de grandeza dos períodos fundamentais esperados. Para cada quadro analisado, é extraída a freqüência fundamental e registrada sua duração. Se a aplicação do método Yin retornar o mínimo global da FDAAN, determina-se que o tamanho do quadro deve aumentar. O resultado obtido para a freqüência fundamental do quadro é descartado e a análise é realizada novamente, desta vez com um quadro com a nova duração. Se o valor retornado foi um mínimo local, o quadro é deslocado adiante num passo igual à metade de sua duração e então, se foi encontrado um período fundamental menor que um quarto da duração do quadro, este é reduzido à metade. Para cada análise, registra-se a duração do quadro utilizado e a freqüência fundamental obtida, obtendo-se uma seqüência que permite determinar a freqüência fundamental do sinal sonoro a cada instante.

3 PÓS-PROCESSAMENTO

Para o agrupamento correto dos resultados do processo de detecção de freqüência fundamental e a eliminação de eventuais erros, é necessária a aplicação de um algoritmo de pós-processamento.

O algoritmo de pós-processamento proposto inicia-se com a conversão das freqüências fundamentais reconhecidas para o valor MIDI da nota musical correspondente através da Equação 3.

$$N = round\left(12\log_2\left(\frac{f_o}{440}\right)\right) + 69 \qquad (3)$$

A operação *round* corresponde ao arredondamento do valor calculado para o inteiro mais próximo.

Pode-se então agrupar a seqüência de notas de forma que elementos subseqüentes apresentando a mesma nota são substituídos por um elemento relacionado à mesma nota e cuja duração é a soma das durações dos elementos originais. Essa substituição permite mais simplicidade nas análises posteriores.

Para a eliminar eventuais erros do processo transcritivo, é preciso conhecer a origem dos erros mais comuns.

O erro mais comum decorre do ataque, estado transitório presente no começo da execução de uma nota. Nesse período, a onda musical está em processo de estabelecimento no corpo do instrumento, levando o processo de detecção de freqüências fundamentais a apresentar falhas. A partir do conhecimento sobre o tempo de ataque relacionado ao instrumento, nota e técnica utilizada, podemos determinar a duração mínima de notas musicais detectáveis. Esse tempo varia entre dois instrumentos diferentes e pode ser modificado, embora o uso de um tempo fixo de 40ms tenha apresentado bom desempenho para todos os instrumentos testados. Diante desse aspecto, classifica-se como ruído qualquer freqüência fundamental com duração inferior a 40ms.

Embora erros quase sempre sejam decorrentes do ataque, eles podem ocorrer também durante a execução da nota. Quando uma nota de duração menor que o limite inferior estabelecido está entre duas outras notas iguais, sua altura é modificada para a altura das notas adjacentes. Caso contrário, a nota errônea é classificada como um ataque e é substituída por silêncio de mesma duração. Após esse processo, a seqüência obtida é novamente agrupada. São excluídas da nova seqüência obtida as notas restantes cuja duração é insuficiente para apresentar uma sensação de tom [6]. Neste trabalho, a duração adotada foi de 50ms.

4 RESULTADOS

Na etapa de testes, o algoritmo de transcrição com variação do tamanho do quadro foi implementado em MATLAB, adicionando-se limites superior e inferior para o tamanho do quadro em 2048 e 256 amostras, respectivamente. Seu desempenho foi comparado ao de outros transcritores baseados no método Yin, mesmo algoritmo de pós-processamento e quadros de duração fixa em 256, 512, 1024 e 2048 amostras. Todos os sinais analisados nessa etapa foram amostrados a 44,1kHz, e estão disponíveis na base de dados de arquivos de som da universidade de Iowa [7]. Essa base de dados contém notas, tocadas por diversos instrumentos, em seqüência crescente de semitons e com tempo de silêncio significativo entre cada uma delas. Essa característica a torna ideal para uma primeira verificação de erros devidos à atuação sobre diferentes tons e timbres.

Os resultados, expressos na Tabela 1, foram classificados em notas corretas (cuja nota estimada corresponde à nota executada em cada intervalo de tempo), perdidas (notas que foram perdidas por conseqüência de uma duração inadequada do quadro), erradas (notas que não foram detectadas por ou que foram detectadas de forma errada), e, em adição a essa estatística, foram adicionadas as notas fantasmas, aquelas que o algoritmo transcritor detecta embora não existam realmente. A freqüência de notas fantasmas (F.N.F.) foi computada através divisão do número de notas fantasmas pelo número de notas corretas obtidas.

Considerando o índice de notas corretamente detectadas, os resultados mostram que o transcritor baseado em quadros de duração variável tem desempenho semelhante ao desempenho do transcritor com quadros de duração fixa em 2048 amostras. Outro aspecto a ser levado em consideração é que, com a variação do tamanho da quadro, pequenos erros durante o processo de transcrição de notas mais agudas se limitaram a intervalos de tempo proporcionais ao período fundamental da nota avaliada, podendo, por conseqüência, ser facilmente removidos. Ao se fixar o quadro em um tamanho suficientemente grande para abranger toda a gama de freqüências da música – 2048 amostras – pequenas flutuações na execução de notas agudas tornam-se mais significativas, resultando no aparecimento de mais notas fantasmas.

Tabela 1: Resultados da transcrição aplicada à base de dados da Universidade de Iowa

Quadro	**Var.**	**256**	**512**	**1024**	**2048**
Notas avaliadas	1163	1163	1163	1163	1163
% correta	88,91	43,59	69,91	81,94	89,25
% perdida	0,60	49,44	22,18	8,68	0,60
% errada	10,49	6,96	7,91	9,37	10,15
F.N.F.	8,03	7,69	10,70	12,49	13,97

Por tratar-se de uma base de dados extensa, o tempo de execução também foi registrado, em minutos, de forma a

verificar o ganho real de eficiência. Os algoritmos rodaram em um computador Athlon XP 2200+ com 236Mb de memória RAM. Os tempos de execução, em minutos, estão mostrados na Tabela 2. Verifica-se que o transcritor com quadros de duração variável apresentou eficiência substancialmente maior que o transcritor com quadros de tamanho fixo em 2048 amostras.

Tabela 2: Tempo de execução durante o teste

Quadro	**Var.**	**256**	**512**	**1024**	**2048**
Tempo (min)	78,94	43,70	49,83	69,88	107,35

Na segunda etapa de testes, algumas melodias foram submetidas aos transcritores de quadro variável e de quadro fixo em 2048 amostras. As melodias foram compostas de trechos de dez a vinte segundos de músicas conhecidas (versão do Bolero de Ravel, executada em um trompete, Asa Branca e Greensleeves, executadas em uma flauta doce, e Brasileirinho, executado em um violão), em gravações com considerável ruído de captação. Todas os sinais foram amostrados a 44,1kHz. As notas foram classificadas em corretas (cuja duração e nota, pela Equação 3, foram estimadas corretamente), perdidas (que não foram encontradas pelo sistema) e com a duração errada (que a altura foi estimada corretamente, mas não a altura). Foram também contabilizadas as notas fantasmas. Os resultados estão expressos na Tabela 3.

Tabela 3: Desempenho na transcrição de melodias gravadas

Quadro	**Variável**	**2048**
Total de notas avaliadas	88	88
Total de notas corretas	82	85
Duração errada	5	2
Notas perdidas	1	1
Notas fantasmas	1	12

Como pode ser verificado, o desempenho de um transcritor baseado no algoritmo proposto é semelhante ao de um transcritor com duração de quadro fixa em 2048 amostras, apresentando, porém, tempo de execução 30% menor.

5 CONCLUSÃO

Este artigo abordou o problema da transcrição automática em áudio monofônico. Foi apresentado um algoritmo para variação da duração do quadro, com o objetivo de otimizar o compromisso entre o limite inferior de freqüência detectável e a resolução temporal do processo.

Os resultados alcançados mostram que a variação da duração do quadro pode levar a ganhos significativos para o algoritmo de transcrição, não só no desempenho geral como no esforço computacional demandado.

Futuramente, pretende-se estudar a aplicabilidade da variação da duração dos quadros em algoritmos mais complexos, como transcritores para áudio polifônico.

6 REFERÊNCIAS

[1] N. Trevillato, J. G. A. Barbedo, A. Lopes, "Transcrição automática de sinais de áudio", anais do X Simpósio Brasileiro de Computação Musical, pp. 291-294, 2005.

[2] L. P. Clarisse, *et all*, "An auditory model based transcriber for singing sequences" , Proc of 3rd International Conference on Music Information Retrieval, ISMIR, '02, 2002.

[3] D. Gerhard, "Pitch extraction and fundamental frequency: history and current techniques" Technical report TR-CS 2003-06, University of Regina, 2003.

[4] F. Hamidi-Toosi, J. Laska "Pitch Detection for the Next Millenium and Beyond", www.ews.uiuc.edu/~laska/ece320/files/Report.pdf, 2004.

[5] A. Cheveigné, H. Kawahara, "YIN, a fundamental frequency estimator for speech and music" J. Acoust. Soc. Am. Vol. 111 No. 4, pp. 1917-1930, 2002.

[6] H. F. Olson, "Music, Physics and Engineering" ed. Dover, 1967.

[7] MIS – Musical Instrument Samples, http://theremin.music.uiowa.edu/MIS.html .