Segmentação de Canto dentro de Sinais de Música

Phabio J. Setubal, Sidnei Noceti Filho e Rui Seara

***Resumo*—Este trabalho propõe uma abordagem para realizar a segmentação automática do canto dentro de sinais de música, baseado na diferença entre o conteúdo harmônico dinâmico dos sinais de canto e dos instrumentos musicais. Os resultados obtidos são comparados com os de outra abordagem proposta na literatura. Obtém-se para ambas uma taxa de acerto em torno de 80%, mesmo considerando um método de medida de desempenho mais rigoroso na avaliação de nossa proposta. Como vantagem, nossa abordagem apresenta menor complexidade computacional. Adicionalmente, são discutidos resultados referentes aos tipos de erro envolvidos no processo de segmentação.**

***Palavras-Chave*—Conteúdo harmônico dinâmico, *pitch*, segmentação de canto, trilhas de picos espectrais, vibrato.**

Abstract—This paper proposes an approach to accomplish the automatic segmentation of singing voice within music signals, based on the difference between the dynamic harmonic content of singing voice signals and that of musical instruments. The results obtained are compared with the one of another approach proposed in the literature. It is obtained for both an accuracy rate around 80%, even using a more rigorous performance measure for our approach. As an advantage, the new procedure presents lower computational complexity. In addition, we discuss results referring to the error types involved in the segmentation process.

***Index Terms*—Dynamic harmonic content, pitch, segmentation of singing voice, spectral peak tracks, vibrato.**

I. INTRODUÇÃO

Atualmente, com a facilidade de acesso a arquivos compactados via *internet* (mp3 e similares), tem-se a possibilidade de se dispor de extensas bibliotecas de áudio de música. Entretanto, uma quantidade limitada de informações pode ser extraída de tais arquivos como, por exemplo, o nome da música, do artista, título do álbum. Essas informações, conhecidas como etiquetas ID3 [1], são ainda inseridas de forma manual e baseadas em texto. A extração automática de outras informações do sinal de música facilitaria a organização da biblioteca. Além disso, tornaria possível imaginar uma série de novas aplicações. Nesse sentido, a segmentação de canto propõe a localização automática dos segmentos contendo canto dentro de um sinal de música. O canto, normalmente, exerce uma função fundamental em uma música, ao carregar as informações da letra e da melodia.

Phabio J. Setubal, Sidnei Noceti Filho e Rui Seara, LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, SC, E-mails: {phabio, sidnei, seara}@linse.ufsc.br.

Este trabalho foi parcialmente financiado pelo CNPq.

Como aplicação direta, a segmentação de canto realiza a indexação em sinais de música, permitindo o acesso direto a trechos contendo ou não canto, facilitando a organização em um banco de dados. Além disso, fornece suporte para outras aplicações, tais como:

i) transcrição automática da letra de uma música, como sugerido em [2];

ii) separação entre canto e instrumentos musicais dentro de sinais de música;

iii) identificação do cantor em uma música [3];

iv) classificação completa de uma peça musical, como sugerido em [4]. Trechos com canto podem ser classificados como canto solo ou em coro. Trechos com canto solo podem ser classificados como voz masculina ou feminina. Trechos instrumentais podem ser classificados pelos tipos de instrumentos executados.

Nestas aplicações, a segmentação de canto constitui uma etapa de pré-processamento. Essa etapa é importante, pois, além de filtrar a informação necessária à etapa seguinte do processo, reduzindo seu custo de processamento, também leva a uma redução na taxa de erro da próxima etapa, visto que são eliminadas informações desnecessárias.

A abordagem proposta neste trabalho é baseada na modificação da técnica apresentada em [5]. Tal técnica sugere a segmentação de um sinal de audiovisual em diversas classes de sinais de áudio, incluindo canto e música puramente instrumental. Apesar da técnica discutida em [5] considerar um processo de segmentação, ela apresenta resultados baseados apenas em experimentos de discriminação. Assim, uma comparação direta de resultados entre a nossa abordagem e aquela apresentada em [5] fica prejudicada. No processo de discriminação, o sinal de música já se encontra pré-segmentado (em trechos com canto e puramente instrumental), bastando classificá-lo corretamente. Por outro lado, em um processo de segmentação, o sinal analisado inclui tanto trechos com canto quanto trechos puramente instrumentais, impondo ao algoritmo de processamento apontar precisamente cada um desses segmentos.

Em [2], é discutido um processo de segmentação com objetivos idênticos aos propostos neste trabalho. Portanto, para efeito de comparação e avaliação da nova proposta, é usado o mesmo banco de dados adotado em [2]. A técnica discutida em [2] utiliza a mesma abordagem proposta em [6], a qual aplica um sistema de reconhecimento de fala (baseado em redes neurais, estimando probabilidade *a posteriori*) para realizar a discriminação entre fala e música. Em [2], a justificativa para o uso da mesma abordagem de [6] é a de que, embora os sinais de canto e fala sejam diferentes, ambos

compartilham algumas características, tais como a estrutura de formantes e a transição de fonemas. Portanto, conjectura-se que um modelo acústico treinado em fala poderia responder de forma similar para sinais de canto, e diferente para aqueles de instrumentos musicais.

Resultados comparando a técnica proposta com aquela considerada em [2] são apresentados, indicando que a nova abordagem é competitiva em termos de taxa de acerto, mesmo usando-se um critério mais rigoroso de avaliação; apresenta menor complexidade computacional; e permite discriminar e discutir os resultados referentes aos diferentes tipos de erro de segmentação envolvidos, sugerindo alternativas para minimizá-los, quando possível.

II. Proposta para Segmentação de Canto

Neste trabalho, a proposta para a segmentação de canto dentro de sinais de música baseia-se na diferença do conteúdo harmônico dinâmico dos sinais envolvidos no processo. Para evidenciar essa diferença, são considerados padrões obtidos da extração de um parâmetro principal. É também proposto um parâmetro secundário adicional que auxilia no processo de segmentação do canto. Finalmente, é considerada uma etapa de suavização dos resultados.

A. Parâmetro Principal: Trilhas de Picos Espectrais

Geralmente em música, tanto o canto quanto os instrumentos musicais possuem características harmônicas. Entretanto, analisando o espectrograma de tais sinais, é possível perceber uma diferença no conteúdo harmônico dinâmico de ambos. Comumente a nota emitida por um instrumento musical é fixa. Assim, a magnitude do espectro se repete durante o tempo em que é a nota é executada, produzindo trilhas constantes no espectrograma, como ilustrado na Figura 1. O canto vozeado, por outro lado, tende a produzir trilhas variáveis no espectrograma, como mostrado na Figura 2. O parâmetro que extrai essa característica harmônica dinâmica em sinais de áudio é proposto em [5], e denominado trilha de pico espectral. Em [5], a identificação de canto é obtida apenas das trilhas variáveis em forma de ondas (vibrato), como ilustrado na Figura 2.

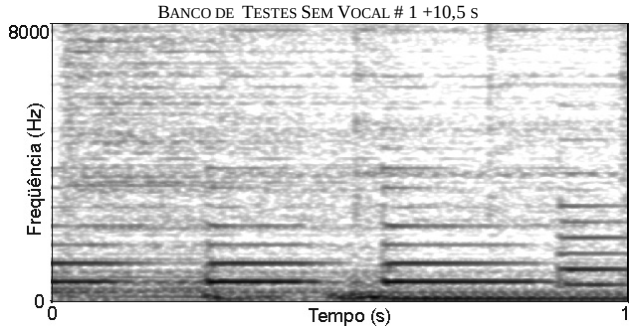

Fig. 1. Espectrograma de um sinal de música puramente instrumental.

Apesar de importante na identificação do canto, o vibrato não está presente em todas as músicas. Além do mais, se presente, é quase certo que não é encontrado em todos os trechos cantados. Contudo, mesmo na ausência do vibrato, a forma das trilhas em segmentos com canto tende a ser variável. A transição dos fonemas de uma palavra durante o canto provoca pequenas variações no valor do *pitch*, como ilustra a Figura 3. Portanto, é sugerida uma alteração na seleção das trilhas de picos espectrais variáveis proposta em [5], de modo a capturar também essas pequenas variações de *pitch*. Recentemente, proposta similar é sugerida em [4], sem, no entanto, ser discutida sua implementação.

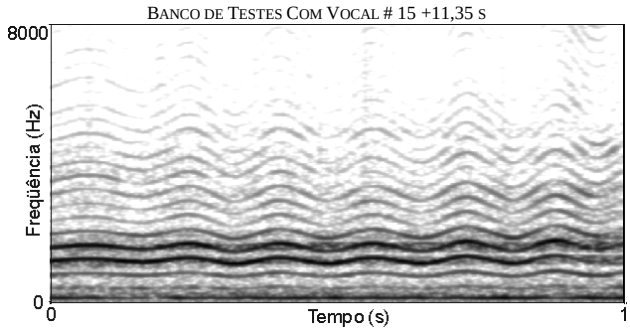

Fig. 2. Espectrograma de um sinal de canto (vibrato).

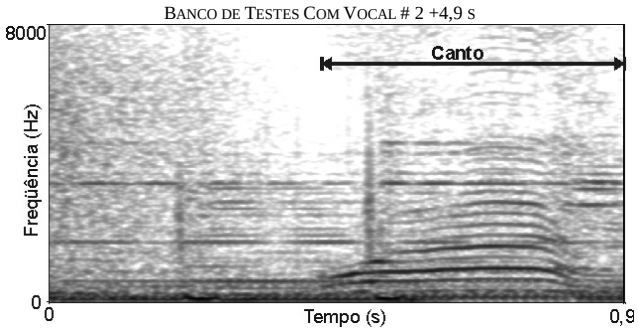

Fig. 3. Espectrograma de um sinal de música com canto caracterizado por uma pequena variação de *pitch*.

Como apresentado em [5], as trilhas de picos espectrais são extraídas a partir de um modelo auto-regressivo (AR), cuja formulação, composta somente por pólos, privilegia a localização de picos espectrais. Em [5], destaca-se que um modelo AR de ordem 40 seria suficiente para extrair os picos espectrais da maioria dos trechos contendo canto. Entretanto, verifica-se que uma ordem 40 não é suficiente para a identificação de canto quando se tem pequenas variações de *pitch*. Assim, adotamos um modelo AR de ordem 80. A Figura 4 apresenta o resultado da captura de todos os picos espectrais do espectrograma da Figura 3, através de modelos AR de ordens 40 e 80. É possível observar que uma ordem 40 não é suficiente para capturar satisfatoriamente as trilhas de canto caracterizadas por pequenas variações de *pitch*. A extração de picos do espectro é realizada a cada 10 ms e o tempo de duração do quadro é de 25 ms com janela de Hamming. A taxa de amostragem usada é de 16 kHz.

Em [5], uma primeira etapa para a extração das trilhas de picos espectrais consiste em excluir alguns picos por esses apresentarem um reduzido valor de amplitude, serem muito suaves, e não estarem relacionados harmonicamente com um

valor de freqüência comum (freqüência fundamental). Em nosso trabalho consideramos a dificuldade de obter, algumas vezes, um valor predominante de freqüência fundamental analisando apenas um quadro. Portanto, prefere-se extrair o máximo de picos nessa etapa, baseado somente na análise de amplitude e suavidade. Assim, opta-se por realizar a análise harmônica por último, após a seleção das trilhas variáveis.

Em uma segunda etapa, como sugerido em [5], os picos espectrais são alinhados e ordenados temporalmente para formar as trilhas de picos espectrais.

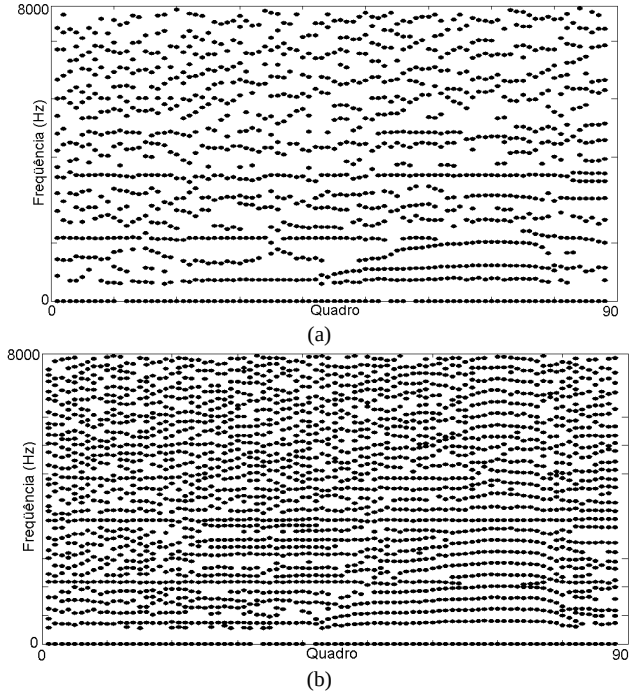

Fig. 4. Captura dos picos do espectrograma da Figura 3 obtidos via modelo AR. (a) Ordem 40. (b) Ordem 80.

Uma vez obtidas as trilhas de picos espectrais ordenadas temporalmente, busca-se identificar os padrões que caracterizam o canto, ou seja, selecionam-se as trilhas variáveis sob a forma de vibrato ou pequenas variações de *pitch*.

Em [7], é observado que a taxa de variação das trilhas resultante do vibrato situa-se, geralmente, na faixa de 4 a 8 Hz. É verificado também que uma taxa de variação de 1 a 3 Hz é mais provável de ocorrer na transição entre fonemas, caracterizando uma pequena variação de *pitch*. Finalmente, é ainda observado que taxas de variação maiores do que um limite superior de 15 Hz estão além do limiar audível e são muito difíceis de serem produzidas pelo trato vocal humano. Portanto, aqui também são consideradas tais restrições. A análise da forma das trilhas é baseada em um critério de diferenças. A medida da taxa de variação para uma determinada trilha, depende da mudança de sinal no valor da diferença entre quadros consecutivos. É considerada tanto a forma completa de uma trilha quanto segmentos parciais. Assim, uma trilha pode ser selecionada ou excluída (completa ou parcialmente), como mostrado pelo exemplo da Figura 5.

Pequenas variações de amplitude de trilhas [ver Figura 5(a), $l \leq 8\,\mathrm{Hz}$], ainda são consideradas trilhas constantes e, portanto, também são descartadas. Trilhas caracterizadas pelo vibrato são, geralmente, selecionadas em sua forma completa, como ilustrado na Figura 5(b). A Figura 5(c) apresenta o exemplo de uma seleção parcial, caracterizando uma pequena variação de *pitch*. Nesse caso, o primeiro trecho é descartado por ser constante, e o segundo, por apresentar uma taxa de variação maior do que 15 Hz.

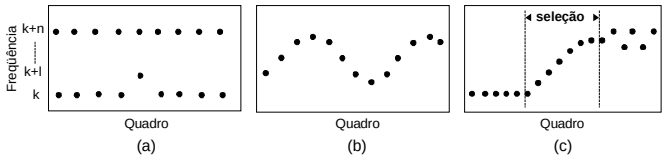

Fig. 5. Análise da forma das trilhas (esboço). (a) Descarte completo. (b) Seleção completa. (c) Seleção parcial.

A seleção de uma simples trilha pela sua forma variável é necessária mas não suficiente para a identificação do segmento correspondente como sendo canto. Como o canto possui característica harmônica, a forma variável de uma trilha deve se repetir para múltiplos inteiros da freqüência fundamental. Assim, para validar um segmento de canto, é necessário obter um mínimo de 3 trilhas harmonicamente relacionadas durante um tempo de duração mínima equivalente a 4 quadros. Essa etapa, de análise harmônica, é a última do processo de seleção de trilhas variáveis.

A Figura 6 mostra o resultado da seleção automática das trilhas de picos espectrais variáveis, aplicado ao sinal da Figura 3. Comparando-se o resultado obtido com aquele decorrente da marcação manual (ver Figura 3), verifica-se que a seleção de trilhas variáveis não identifica corretamente todo o segmento de canto. Dessa forma, a extração de um parâmetro secundário auxiliaria para a redução de tal erro, sendo discutido na próxima seção.

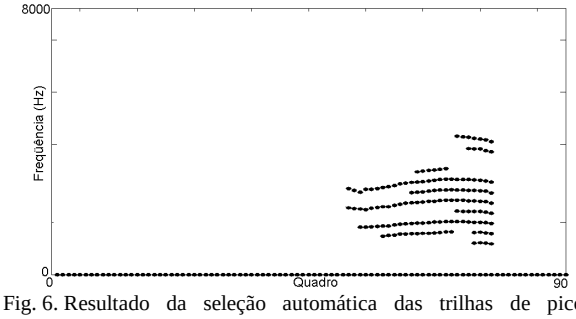

Fig. 6. Resultado da seleção automática das trilhas de picos espectrais variáveis.

B. Parâmetro Secundário: Média da Magnitude

A extração do parâmetro denominado média da magnitude (MM) é obtida pelo cálculo do valor médio da magnitude da transformada de Fourier (TF), a cada quadro do sinal de entrada. Seu valor é calculado para duas bandas de freqüência. Assim,

$$\mathrm{MMB}_q = \frac{\sum_{n=k_1}^{k_2} |X(k)|}{k_2 - k_1}, \qquad (1)$$

onde $|X(k)|$ é a magnitude da TF do sinal, MMB_q é o valor da MM em baixas freqüências para o quadro *q*, calculada para a banda de freqüência entre k_1 e k_2, com $k_2 > k_1$.

O limite inferior $k_1 = 500\,\mathrm{Hz}$ evita a influência de ruídos de baixa freqüência. Como discutido em [5], esse mesmo limite também é considerado na extração das trilhas de picos espectrais. O limite superior $k_2 = 4\,\mathrm{kHz}$ considera que a maior parte da energia do canto vozeado está localizada até esse limite de freqüência.

Para altas freqüências, o valor da MM é calculado por:

$$\mathrm{MMA}_q = \frac{\sum_{n=k_3}^{K} |X(k)|}{K - k_3}, \qquad (2)$$

onde MMA_q é o valor da MM em altas freqüências para o quadro *q*, calculada para a banda de freqüência entre k_3 e K, com $K > k_3$.

O limite superior $K = 8\,\mathrm{kHz}$ equivale ao limite teórico da largura de banda para uma taxa de amostragem de 16 kHz. O limite inferior $k_3 = 5\,\mathrm{kHz}$ é definido como o início da banda de altas freqüências.

Os valores da MMA e MMB são normalizados pelos seus respectivos valores máximos, considerando-se todos os quadros do sinal analisado.

Conforme descrito a seguir, o parâmetro secundário MM auxilia no processamento de segmentação do canto, ajustando os limites do canto não identificados corretamente e detectando padrões secundários característicos do canto, como aqueles destacados na Figura 8(a).

• Ajuste de Limites do Canto Vozeado

Algumas vezes, a seleção automática das trilhas variáveis não identifica corretamente os limites de início e final dos segmentos contendo canto, como observado na Figura 6. Assim, é proposto que a variação de valor da MMB seja usada para estender tais limites. Isso é justificado pelo fato de que, geralmente, o início e o final de um segmento contendo canto vozeado tendem a produzir uma variação maior do valor da MMB. Essa característica é explorada segundo as regras apresentadas a seguir:

i) Analisando toda a extensão do sinal, são escolhidos quadros candidatos a iniciar um segmento de canto, de acordo com valores predefinidos de variações positivas da MMB;

ii) A partir do quadro onde se inicia uma trilha variável, é realizada uma varredura em direção ao início do sinal, buscando encontrar o primeiro quadro escolhido por (i). Esse valor é considerado o novo início do segmento de canto;

iii) O mesmo procedimento é adotado para identificar o final de um trecho com canto. Nesse caso, a varredura segue em direção ao final do sinal, buscando encontrar o primeiro candidato a finalizar o segmento contendo canto, baseado em variações negativas da MMB.

A Figura 7 mostra o resultado do ajuste dos limites do canto vozeado identificados anteriormente pela seleção das trilhas (ver Figura 6). Nota-se que estes novos limites obtidos automaticamente aproximam-se daqueles obtidos através da marcação manual, mostrados na Figura 3. Dessa forma, o erro na segmentação do canto é reduzido consideravelmente.

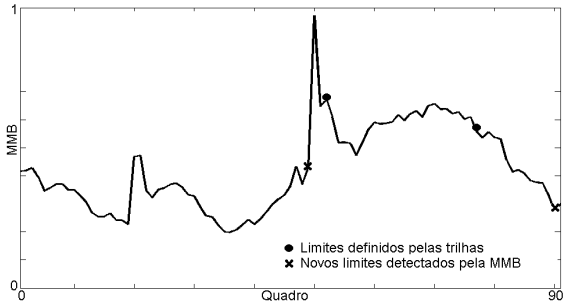

Fig. 7: Ajuste dos limites do canto vozeado pela variação da MMB.

• Detecção do canto fricativo

Segmentos fricativos presentes em trechos com canto não possuem característica harmônica. Assim, não é esperada a formação de trilhas de picos espectrais no domínio tempo-freqüência. Em um trecho de música que contém canto fricativo, havendo a presença de trilhas, certamente essas são devidas a instrumentos musicais.

Os segmentos fricativos do canto possuem uma maior magnitude em altas freqüências, como ilustrado na Figura 8(a). Esse padrão é detectado pelo maior valor da MMA em relação à MMB, respeitando limites mínimos dessa relação e tempos mínimos de duração, como mostrado na Figura 8(b).

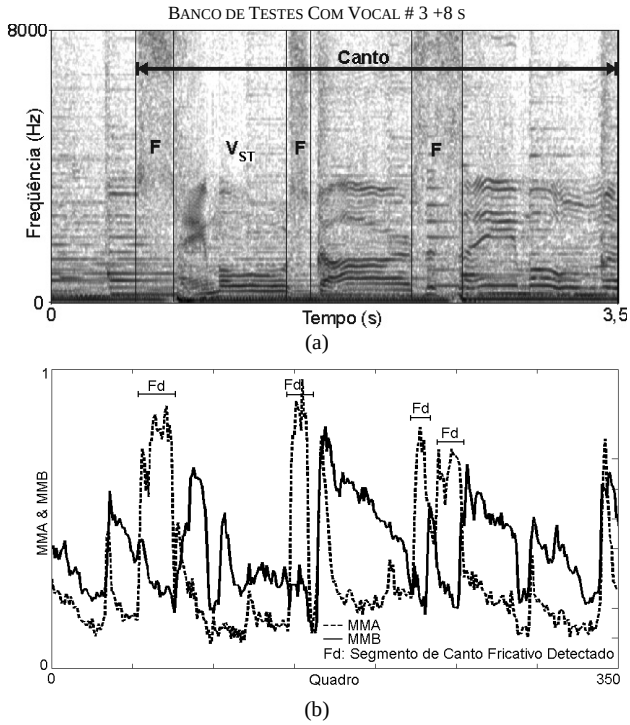

Fig. 8. (a) Espectrograma de um sinal de música destacando segmentos de canto fricativo, denotado por "F", e um segmento de canto vozeado que não produz trilhas variáveis, denotado por "V_{ST}". (b) Detecção dos segmentos de canto fricativo pela relação entre MMA e MMB.

C. *Suavização dos Resultados*

A suavização dos resultados constitui uma etapa de pós-processamento e tem a função de alterar o estado de um trecho intermediário, baseado na predominância de um outro estado em trechos vizinhos. Tal procedimento é aplicado à localização de segmentos intermediários de canto não identificados por trilhas variáveis ou por canto fricativo, como aquele destacado na Figura 8(a), denominado V_{ST}. Assim, respeitando-se certos limites predefinidos, tais como tempo de duração máxima do segmento e valores médios da MMA ou MMB, esse trecho intermediário tem seu estado alterado, dando origem a um único segmento de canto, composto pela seqüência F-V_{ST}-F. Na verdade, neste exemplo da Figura 8(a), os outros segmentos de canto ainda são identificados pela detecção de canto fricativo e pela seleção de trilhas variáveis. Assim, na saída do sistema proposto, obtém-se automaticamente a identificação correta de todo o segmento de canto destacado, obtido previamente usando marcação manual.

No banco de dados usado, existem arquivos de sinais de música puramente instrumental. Portanto, são também aplicadas regras de suavização específicas para esses casos. Assim, se após o processamento automático de todo o sinal de música de entrada, o número e a duração dos trechos identificados como canto forem menores do que limites predefinidos, todos esses trechos são alterados, passando a considerá-los como não contendo canto, e conclui-se que o sinal de música de entrada é puramente instrumental.

III. RESULTADOS

Para avaliação do algoritmo proposto, são utilizados 101 arquivos de música que integram o banco de dados proposto por [8], e também usado em [2], [6]. Cada arquivo tem 15 s de duração, amostrados originalmente a uma taxa de $22,05$ kHz, e gravados aleatoriamente de uma rádio FM em 1996. Em [2], dos 101 arquivos, 61 são usados como banco de treinamento e 40 como banco de testes. A melhor taxa de acerto é obtida para o erro calculado com janelas de 1,3 s. Dentro desse intervalo, não importam os tempos exatos da localização dos trechos com canto, mas sim o tempo total de canto detectado.

Em nosso trabalho, é proposto um método de análise de desempenho mais rigoroso. Na saída do algoritmo obtêm-se os tempos limites dos segmentos contendo canto, com precisão da ordem de centésimos de segundo. Esses tempos são diretamente comparados com os tempos obtidos usando marcação manual. A Tabela 1 compara o desempenho de ambas as técnicas, aplicadas ao mesmo banco de dados de testes, considerando seus respectivos métodos de análise.

TABELA 1
COMPARAÇÃO DO DESEMPENHO ENTRE O ALGORITMO DA REFERÊNCIA [2] E O ALGORITMO PROPOSTO

Algoritmo	*Taxa de Acerto* (%)
Referência [2]	**81,2**
Proposto	**81,7**

Em [2], os limites que distinguem as classes de canto e instrumentos musicais são obtidos através de um classificador automático considerando o banco de treinamento. Em nosso caso, os limites são obtidos de forma manual, através de um procedimento baseado em regras, considerando a análise de uns poucos arquivos de sinais do banco de dados. Portanto, não há necessidade de uma etapa de treinamento. Assim, é permitida a extensão da análise de desempenho do algoritmo para os 61 arquivos desse banco, aumentando a abrangência dos testes com relação à [2]. Como mostrado na Tabela 2, o mesmo nível de taxa de acerto é mantido.

TABELA 2
ANÁLISE DE DESEMPENHO DO ALGORITMO PROPOSTO APLICADO AO BANCO DE DADOS DE TREINAMENTO E AO BANCO TOTAL

Banco de Dados	*Taxa de Acerto* (%)
Treinamento (61 sinais)	**82,6**
Total (101 sinais)	**82,2**

Diferentemente do método de análise de desempenho proposto em [2], nosso método permite discriminar os tipos de erro envolvidos: erro na detecção dos limites inferior ou superior do canto $(E_L = E_{LI} + E_{LS})$; erro ao não detectar um trecho completo de canto, chamado de falso negativo (E_{FN}); erro ao detectar um trecho completo de canto não existente, chamado de falso positivo (E_{FP}). A Tabela 3 apresenta a distribuição da taxa de erro de acordo com o seu tipo.

TABELA 3
DISTRIBUIÇÃO DA TAXA DE ERRO CONFORME OS SEUS TIPOS

Tipo de Erro	*Taxa de Erro* (%)
E_L	**7,7**
E_{FN}	**4,4**
E_{FP}	**5,7**
Total	**17,8**

Nós também sugerimos a adoção de um nível de confiabilidade binária aos sinais de entrada, definido como uma medida de certeza de que o resultado produzido pelo algoritmo esteja correto, baseado em atender um conteúdo harmônico mínimo. Assim, analisando todo o banco de dados, determinou-se que para os sinais que não alcancem, em algum trecho nos 15 s de duração, um valor mínimo de 7 trilhas variáveis harmonicamente relacionadas, é atribuído um nível de confiabilidade igual a zero. Por exemplo, no segmento de canto identificado na Figura 6, obtém-se um trecho com até 9 trilhas harmonicamente relacionadas, garantindo um nível de confiabilidade 1 (um). Excluem-se dessa análise os sinais previamente classificados como puramente instrumentais.

De acordo com o nível de confiabilidade definido, sugere-se realizar a análise de desempenho excluindo os sinais do banco de dados com nível de confiabilidade zero. Nesse caso, comprova-se que o desempenho do algoritmo tende a aumentar, conforme mostrado na Tabela 4.

TABELA 4
DESEMPENHO DO ALGORITMO APLICADO SOMENTE AOS SINAIS COM NÍVEL DE CONFIABILIDADE 1 (UM)

Banco de Dados	*Taxa de Acerto* (%)
Testes (32 sinais)	**86,6**
Total (81 sinais)	**86,9**

IV. DISCUSSÃO

Apesar de propormos um método de análise de desempenho mais rigoroso do que o proposto em [2], constata-se através da Tabela 1 que o desempenho obtido em ambos os casos é similar, na faixa de 80%. Além do mais, uma vez que nossa abordagem considera um procedimento baseado em regras, ela apresenta um menor esforço computacional quando comparada à proposta de [2], que necessita inclusive de um sistema de reconhecimento de fala.

Como anteriormente mencionado, nossa proposta ainda permitiu a extensão da análise de desempenho para todos os 101 arquivos de sinais do banco de dados, mantendo o mesmo nível da taxa de acerto, conforme mostrado na Tabela 2.

Outra vantagem, com respeito à técnica proposta em [2], é a possibilidade de discriminar os tipos de erros envolvidos (ver Tabela 3). Através da análise desses diferentes tipos de erro, é possível adotar algumas alternativas para reduzi-los, quando possível.

Basicamente, dois motivos principais contribuem para a ocorrência de erros na segmentação de canto a partir da abordagem proposta: seleção de trilhas variáveis de forma insatisfatória (sinais com baixo conteúdo harmônico), e seleção de trilhas variáveis produzidas por instrumentos musicais.

Para melhorar o desempenho no caso da seleção de trilhas de forma insatisfatória, é adotado um nível de confiabilidade aos sinais de entrada, com base no atendimento de um conteúdo harmônico mínimo. Em relação ao desempenho obtido na Tabela 2, verifica-se através da Tabela 4, um ganho de quase 5% quando o algoritmo é aplicado somente aos sinais com confiabilidade um. Grande parte desse ganho deve-se à eliminação do erro do tipo E_{FN}. Aos demais sinais (confiabilidade zero), é sugerida a aplicação de outras técnicas associadas à abordagem proposta, as quais não considerem o conteúdo harmônico dinâmico dos sinais de entrada.

Os erros tipo E_{FP}, geralmente, devem-se a instrumentos musicais que produzem trilhas variáveis. Particularmente, no banco de dados analisado, alguns tipos de instrumentos (com predominância do saxofone) contribuem com a maior parte da taxa de erro de E_{FP} (ver Tabela 3). Nesse caso, sugere-se a adição de técnicas de reconhecimento de instrumentos, como a proposta em [9].

Os tipos de erros mais grosseiros são os E_{FN} e E_{FP}. Nesses casos, o resultado obtido é oposto à situação real. Entretanto, como visto, é possível investigar algumas alternativas para reduzi-los. Por outro lado, analisando a distribuição da taxa de erro conforme os seus tipos (ver Tabela 3), verifica-se que o erro de maior ocorrência é do tipo E_L. A detecção automática dos limites do canto pode ser considerada como o ponto mais crítico, especialmente porque há um maior grau de subjetividade nas suas definições, e não é possível utilizar a suavização dos resultados, realizada somente para trechos intermediários. Entretanto, apesar de maiores dificuldades em sua detecção, do ponto de vista da compreensão da informação segmentada, os resultados não são tão críticos. Lembre-se que o valor de E_L ainda deve ser distribuído entre o erro na detecção dos limites superior e inferior. Assim, geralmente, na audição do trecho segmentado, o erro na detecção dos limites não compromete a compreensão do conteúdo correspondente.

V. CONCLUSÕES

Este trabalho focou o problema de segmentação do canto dentro de sinais de música, baseado, sobretudo, na diferença entre o conteúdo harmônico dinâmico do canto e dos instrumentos musicais. Assim, propõe-se a extensão dos padrões identificados por um parâmetro consagrado na literatura, e a introdução de um novo. O desempenho do algoritmo proposto foi comparado com o apresentado em [2], utilizando o mesmo banco de dados. A taxa de acerto obtida em ambos situa-se na faixa de 80%. Entretanto, nossa abordagem, além de considerar um critério de medida desempenho mais rigoroso, apresenta uma menor complexidade computacional. Adicionalmente, novos resultados obtidos a partir da extensão dos testes para todo o banco de dados e da discriminação dos tipos de erro são discutidos. A partir da análise dos tipos de erro, são sugeridas alternativas de reduzi-los como, por exemplo, a inclusão de outras técnicas e a adoção de um nível de confiabilidade aos arquivos de sinais de entrada.

AGRADECIMENTOS

Gostaríamos de agradecer ao doutorando Adam Berenzweig, ao Prof. Dr. Daniel Ellis, e aos pesquisadores Dr. Eric Sheirer e Dr. Malcom Slaney, por disponibilizarem o banco de dados de sinais de música que foi usado neste trabalho.

REFERÊNCIAS

[1] M. Nilsson, "ID3v2". Last access in Oct. 2003. [Online]. Available: http://www.id3.org.

[2] A. L. Berenzweig and D. P. Ellis, "Locating Singing Voice Segments within Music Signals," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 119-122, Oct. 2001.

[3] T. Zhang, "Automatic Singer Identification," *Proc. International Conference on Multimedia and Expo*, vol. 1, pp. 33-36, Jul. 2003.

[4] T. Zhang, "Semi-Automatic Approach for Music Classification," *SPIE Conference on Internet Multimedia Management Systems IV*, Sept. 2003.

[5] T. Zhang and C.-C. J. Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 441-457, May 2001.

[6] G. Williams and D. Ellis, "Speech/Music Discrimination Based on Posterior Probability Features," *Proc. EUROSPEECH*, pp. 687-690, Sept. 1999.

[7] D. B. Gerhard, "Computationally Measurable Temporal Differences between Speech and Song," Burnaby, BC, Canadá, *Tese de Doutorado* (Ciências da Computação), Simon Fraser University, Apr. 2003.

[8] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," Proc. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1331-1334, Apr. 1997.

[9] T. Kitahara, M. Goto, and H. Okuno, "Musical Instrument Identification Based on F0-Dependent Multivariate Normal Distribution," *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 421-424, Apr. 2003.