Sociedade de Engenharia de Áudio
Artigo de Convenção
Apresentado na VII Convenção Nacional
26-28 de maio de 2003, São Paulo, Brasil

Transcrição Musical Automática com Bancos de Filtros

Cristiano N. dos Santos, Luiz W. P. Biscainho, Sergio L. Netto
Universidade Federal do Rio de Janeiro
Rio de Janeiro, 21945-970, RJ, Brasil
{csantos, wagner, sergioln}@lps.ufrj.br

RESUMO

Este trabalho investiga o uso de bancos de filtros FRM-CM como ferramenta auxiliar na transcrição automática de sinais musicais. Esses bancos possibilitam alta seletividade com elevado número de bandas, atingindo a alta resolução na freqüência requerida pela aplicação em questão. O artigo discute o problema da transcrição com suas dificuldades inerentes e compara o desempenho dos FRM-CMFBs com o da DFT na identificação de notas musicais.

I - INTRODUÇÃO

A transcrição de sinais de música [1] consiste em gerar, a partir de um sinal de áudio gravado, a representação da música executada numa forma que permita sua reprodução por um músico. A célula básica dessa representação é a nota musical. Assim, pode-se dizer que o coração de um sistema de transcrição é a identificação da altura, do tempo de início e da duração de cada nota emitida. A realização automática desse processo encontra aplicações que vão do auxílio ao ensino musical ao registro e estudo de interpretações de grande valor histórico-musical.

De modo geral, a identificação das notas musicais é feita a partir das informações espectrais do sinal em análise ao longo do tempo. Bancos de filtros [2] são um exemplo de ferramenta que permite a descrição dinâmica de sinais no domínio da freqüência de forma eficiente. Este trabalho analisa o uso de uma família específica de bancos de filtros no problema da transcrição de sinais musicais. Esses bancos se baseiam nas técnicas de mascaramento da resposta na freqüência (*frequency-response masking*, FRM) e de modulação por cossenos (*cosine-modulated filter bank*, CMFB), e possuem alta seletividade e grande número de bandas [3, 4].

A organização deste trabalho segue a seguinte estrutura: Na Seção II, apresenta-se o problema da transcrição musical, discutindo-se sua abrangência e revisando-se sua abordagem prática. Na Seção III, caracteriza-se o sistema de transcrição adotado neste trabalho. Na Seção IV, apresenta-se brevemente o banco de filtros FRM-CM usado neste trabalho. Na Seção V, compara-se o uso dos FRM-CMFBs com a DFT. Por fim, na Seção VI, apresenta-se um conjunto de experimentos demonstrando o uso dos FRM-CMFBs em diversas situações tópicas em transcrição.

II - O PROBLEMA DA TRANSCRIÇÃO

Uma das áreas mais abrangentes do processamento digital de sinais de áudio é a que se dedica à análise e à síntese de sinais musicais, a qual encontra aplicação na restauração e na remixagem de gravações, na síntese de instrumentos, na transcrição automática etc. De uma forma bem genérica, podemos dizer que os principais modelos utilizados em análise e síntese se enquadram em três categorias: puramente estocásticos (como o autorregressivo [5]), puramente

determinísticos (como o senoidal [6]) e simultaneamente estocásticos e determinísticos [6].

Neste trabalho, aborda-se particularmente o problema de transcrição musical automática [1]. Estritamente falando, a definição de sua meta poderia ser: “A partir de uma gravação musical, gerar uma partitura convencionalmente notada.” Nesses termos, a saída do sistema de transcrição deveria permitir a execução da peça musical por um músico sem nenhum treinamento adicional. Contudo, uma proposta genérica como essa pode atingir uma complexidade intratável, como veremos a seguir.

Deseja-se transcrever a peça musical com que grau de expressividade? Por exemplo, aspectos referentes à dinâmica (variações de intensidade) e à agógica (variações de velocidade) da execução musical podem estar associados à intenção do compositor (como ele as indicaria ao escrever a peça) ou à opção interpretativa do executante. Em alguns instrumentos, também o timbramento pode sofrer modificações essenciais (como no caso das diversas formas de se tanger uma corda de violão). Até que ponto é preciso notar os aspectos interpretativos contidos na gravação-fonte?

Outro aspecto de grande impacto na complexidade da tarefa de transcrição diz respeito aos tipos das fontes sonoras e à forma de encará-las. Por exemplo, pode-se considerar o caso de um instrumento solista; este pode ser monofônico (aqui no sentido de emitir uma única nota por vez) ou polifônico (capaz de emitir acordes). No caso de mais de um instrumento, além da classificação anterior, pode-se tratá-los como indivíduos solistas (como no caso de um quarteto de cordas) ou por famílias (como todos os segundos violinos de uma orquestra tocando em uníssono). A complexidade envolvida no reconhecimento de notas simultâneas emitidas por um mesmo (ou mesmo tipo de) instrumento é seguramente maior que no caso da emissão de uma nota por vez por tipo de instrumento. De qualquer modo, há que se determinar que notas são tocadas, e por quem.

Por último, tratar instrumentos temperados (no sentido de só serem capazes de emitir notas com afinação predeterminada, como um piano) permite simplificações que os demais instrumentos (como um violoncelo, em que se pode realizar um *glissando*) não admitem.

Claro que as soluções de melhor desempenho tendem a ser mais específicas; por outro lado, as soluções mais gerais têm o atrativo de serem mais automáticas. No final, a complexidade do sistema a projetar será determinada pela delimitação das metas a alcançar; vamos, então, dimensionar nossos objetivos e descrever as técnicas tipicamente associadas a eles.

III - METODOLOGIA BÁSICA

Primeiramente, a fim de não entrarmos no mérito da notação musical, assumiremos que se discute um sistema que tem como saída uma representação intermediária entre o sinal e a pauta, em princípio contendo em si todas as informações necessárias para gerá-la, se desejado. Como já foi dito, basta caracterizar as notas componentes do áudio sob análise. Aqui, cabe um comentário: a percepção musical pelo homem não se dá pela decomposição do áudio em suas notas individuais, mas a partir da cognição de entidades bem mais complexas e suas interrelações. Entretanto, se o objetivo final da transcrição é gerar uma representação segundo a notação musical convencional, a questão perceptiva perde bastante de sua importância na busca de soluções para este problema.

Outra simplificação que não prejudica tanto a generalidade da discussão é supor que os sinais são predominantemente "tonais" (mais uma vez, uma liberdade de terminologia): referimo-nos à presença exclusiva de fontes sonoras que emitem notas definidas, ao menos em regime permanente, excluindo, por exemplo, a maior parte dos instrumentos de percussão. Mais que isso, todos os instrumentos emitiriam sinais harmônicos, ou seja, compostos de uma componente fundamental numa freqüência f_0 (que definiria, afinal, a nota emitida) e componentes parciais em freqüências múltiplas inteiras de f_0.

Nesse contexto, um sistema típico de transcrição se preocuparia em reconhecer e descrever três aspectos: a entrada (*onset*) e duração de cada conjunto de notas, a individualização das notas com suas freqüências componentes e respectivas amplitudes e a identificação das fontes sonoras individuais, se for o caso.

A identificação do ataque de notas ou acordes geralmente é realizada a partir do exame da envoltória do sinal. A presença de picos na envoltória indica o início de novas emissões sonoras pelas fontes. Essa etapa permite extrair informações sobre a estrutura rítmica da música, que terá importância na sua notação final. A prévia separação do sinal em sub-bandas de freqüência permite, evidentemente, melhor desempenho, já que pode detectar melhor entradas com intensidades diferentes em faixas de freqüência distintas.

A parte mais importante do sistema de transcrição é a de descrição acurada do comportamento espectral do sinal no tempo. Isso envolve, basicamente, identificar cada linha espectral presente no sinal (incluindo todas as freqüências fundamentais e harmônicas que o compõem) e suas respectivas intensidades ao longo do tempo. As linhas precisam ser descritas de tal forma que se possa dizer quando nascem e morrem. O comportamento de uma fundamental e suas harmônicas associadas é coerente em regime permanente. Contudo, dependendo dos instrumentos tratados, as linhas podem variar continuamente em amplitude ou freqüência por efeitos como *tremolo* e *vibrato*. Num *glissando*, por exemplo, percorre-se uma série contínua de notas sem interromper a linha emitida. No caso de emissão simultânea de notas, é possível ocorrer o cruzamento e mesmo a superposição continuada de linhas de freqüência. Um dos casos mais complexos a enfrentar é a distinção entre outra fundamental presente na oitava superior de uma nota e sua segunda harmônica. Aqui, tudo se resume à busca de representações adequadas em tempo-freqüência, das quais as mais populares são a DFT e os bancos de filtros.

A DFT (*Discrete Fourier Transform*) fornece a descrição espectral de blocos de sinal, caracterizando módulo e fase de raias espectrais linearmente espaçadas. Associada ao modelo senoidal, que descreve cada linha freqüencial do sinal como uma soma de senóides harmônicas com amplitudes e fases lentamente variáveis, a DFT e suas variantes são a ferramenta preferida para implementar essa etapa dos sistemas de transcrição. Alternativamente, bancos de filtros permitem analisar continuamente no tempo a energia contida nas diversas regiões do espectro. Se os filtros individuais têm faixas de passagem suficientemente estreitas, permitem descrever acuradamente amplitude e freqüência ao longo do tempo para as componentes do sinal.

Dois aspectos ainda merecem ser mencionados: 1) A escala musical atualmente adotada no ocidente emprega temperamento igual, em que o menor intervalo (o semitom) corresponde a uma razão fixa de $2^{\frac{1}{12}}$ entre notas adjacentes. Então, mesmo que nos restrinjamos a instrumentos de afinação fixa, distinguir uma fundamental de outra exige acurácia melhor que 6%. 2) A importância da amplitude relativa das freqüências harmônicas está na identificação de instrumentos, quando for o caso: sua comparação contra um padrão pode fornecer uma pista muito importante.

Enquanto o problema de detecção de notas individuais já parece hoje satisfatoriamente solucionado, o polifônico ainda requer muito trabalho para ser resolvido. É nesse contexto que se insere este trabalho, onde propomos usar bancos de filtros de alta seletividade nessa etapa do processamento.

IV - O BANCO DE FILTROS FRM-CM

Os FRM-CMFBs são uma família de bancos de filtros que podem ser projetados com seletividade muito alta e elevado número de bandas. O detalhamento do projetos dos FRM-CMFBs pode ser encontrado, por exemplo, em [3] e [4].

Neste trabalho, utilizamos um banco de filtros de 1024 bandas com fator de *rolloff* ρ=0,1, atenuação máxima na banda passante A_p=0,2dB e atenuação mínima na banda de rejeição A_r=60dB . Com estas especificações, o projeto de um banco de filtros convencional seria impraticável. A resposta em magnitude deste FRM-CMFB é detalhada na Fig.1.

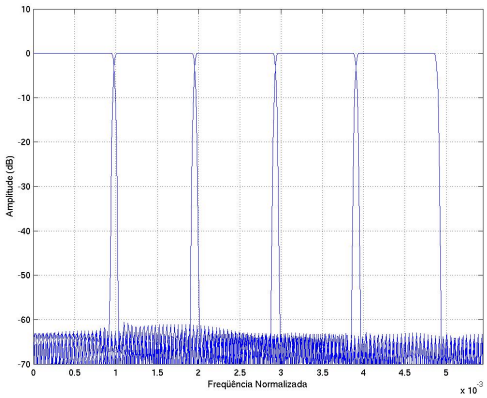

Figura 1: Resposta em magnitude do FRM-CMFB com 1024 em bandas (detalhe das 5 primeiras bandas).

V - FRM-CMFB X DFT

A separação de harmônicos é de fundamental importância na transcrição musical para garantir uma identificação segura das notas e das famílias de intrumentos. Esta separação deve ocorrer em dois níveis: devemos separar os harmônicos uns dos outros e também do ruído presente nas bandas vizinhas do sinal. O uso do FRM-CMFB procura resolver duas dificuldades na separação de harmônicos, comuns na tarefa de transcrição musical.

A primeira dificuldade é a da interferência entre bandas adjacentes. Analisando-se as transformadas mais usadas (DFT, CQT - *Constant-Q Transform* e BQT - *Bounded-Q Transform* [1]) como bancos de filtros, podem-se observar características comuns, como a baixa atenuação das bandas de rejeição de seus filtros. A conseqüência disso é a interferência de informação musical e/ou ruidosa entre bandas adjacentes, podendo tornar difícil a identificação de picos no domínio da freqüência. A Fig. 2 mostra a resposta em magnitude do banco de filtros correspondente à DFT, onde fica evidente a interferência entre bandas adjacentes devido à baixa atenuação (apenas 13 dB) na banda de rejeição dos subfiltros DFT. Comparando-se as Figs. 1 e 2, vê-se que a separação de harmônicos em bandas próximas deverá ser melhor realizada pela estrutura FRM-CMFB.

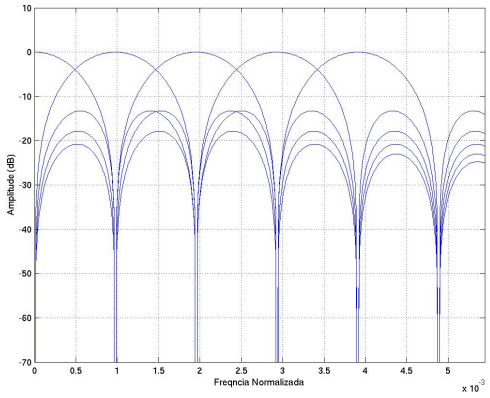

Figura 2: Resposta em magnitude da DFT com 1024 em bandas (detalhe das 5 primeiras bandas).

Em relação à DFT, cabe destacar ainda a presença de um *ripple* significativo na banda passante de seus filtros, o que pode gerar uma atenuação diferenciada das componentes de freqüência, dependendo das posições dos harmônicos em relação ao centro da banda passante. Este aspecto é crítico, por exemplo, na identificação de instrumentos musicais.

A segunda dificuldade diz respeito ao uso de transformadas logarítmicas, como a CQT e a BQT. Estas transformadas se destinam a separar faixas de freqüências a intervalos geométricos. Assim sendo, se é adotada uma resolução de um semitom, para cada nota da escala cromática há uma faixa entre um quarto de tom abaixo e um quarto de tom acima que é separada naquela banda. se a resolução é de quarto de tom, haverá duas faixas a cada nota, e assim por diante. Isso permite ter maior eficiência na distribuição de amostras da transformada, de acordo com a escala ocidental, logarítmica. Porém, essa distribuição tende a agrupar em uma mesma banda os harmônicos de notas diferentes que resultem muito próximos. Conseqüentemente, embora distintos, estes não são identificáveis individualmente.

Para exemplificar esse problema, vamos considerar o caso de um intervalo recorrente na música ocidental, a terça maior. Tomando-se a nota Dó como padrão, sua terça maior superior corresponde ao próximo Mi. Conforme a escala natural, a razão entre suas freqüências deveria ser de 4 para 5, o que levaria o quinto harmônico de Dó a coincidir exatamente com o quarto harmônico de Mi. Podem-se imaginar, em situações de mais notas simultâneas, as ambigüidades bastante complexas que essas superposições podem provocar. Entretanto, na escala de temperamento igual, a terça maior tem um erro de 0,8% em relação à razão de inteiros. Com Dó=262 Hz e Mi=330 Hz, esse desvio faria os dois

harmônicos citados distarem de 10,5 Hz. Uma acurácia suficientemente alta já resolveria a ambigüidade anterior, ao menos para instrumentos de afinação fixa e temperamento igual. As Figs. 3 e 4 mostram, respectivamente, as respostas da BQT, com resolução de um quarto de tom, e do FRM-CMFB, com resolução de 5,4 Hz, a um sinal com freqüências no intervalo musical descrito acima.

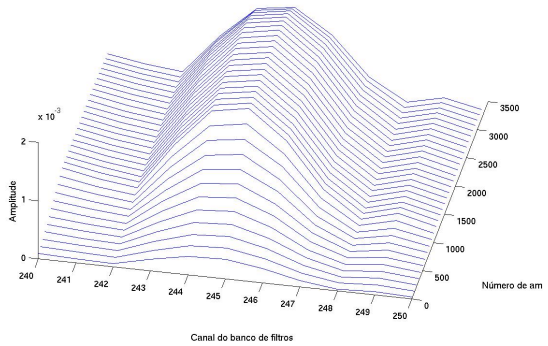

Figura 3: Envoltória da resposta da BQT no intervalo da terça maior.

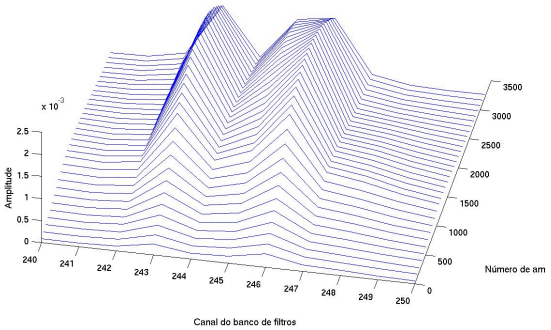

Figura 4: Envoltória da resposta do FRM-CMFB no intervalo da terça maior.

VI - EXEMPLOS DE TRANSCRIÇÃO

Exemplo 1 - Harmônicos Próximos:

O problema de interferência entre bandas pode ser visualizado a partir de um simples exemplo. Considere o caso em que há dois harmônicos situados um na banda *i* e o outro na banda *i*+2, deixando a banda do meio, *i*+1, sem informação relevante. Em caso de baixa atenuação nas bandas adjacentes (como na DFT), os harmônicos nas bandas *i* e *i*+2 podem se mesclar na banda *i*+1, confundindo o processo de transcrição.

Este fenômeno pode ser ilustrado usando-se uma senóide de 7 Hz a mais do centro da primeira banda e outra senóide de 7 Hz a menos do centro da terceira banda. A Fig. 5 mostra a envoltória das saídas de cinco bandas determinadas com a DFT. Note que neste caso não foi possível visualizar os dois harmônicos separadamente devido à baixa atenuação da DFT. O resultado do mesmo experimento usando o FRM-CMFB é visto na Fig. 6, onde podemos claramente perceber a existência dos dois harmônicos.

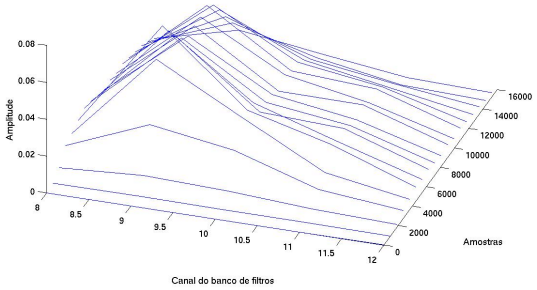

Figura 5: Exemplo 1- Envoltória dos sinais no banco de filtros DFT para o caso de dois harmônicos próximos.

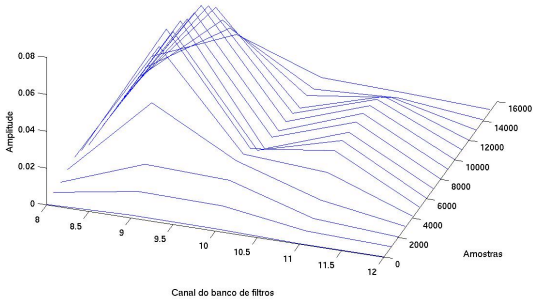

Figura 6: Exemplo 1- Envoltória dos sinais no FRM-CMFB para o caso de dois harmônicos próximos.

Exemplo 2 - Transitórios no Tempo:

Ao se filtrar um sinal do tipo musical, com componentes de freqüência bem definidas, deseja-se obter na saída a(s) componente(s) de freqüência correspondente(s) à banda deste filtro. Mas a filtragem altera a forma de onda, impondo transitórios. Os transitórios podem ser divididos em dois tipos: de subida e de estabilização.

Denominamos transitório de subida aquele em que a amostra central da resposta ao impulso do filtro ainda não chegou à amostra inicial do trecho estacionário. Denominamos de transitório de estabilização aquele em que a amostra central do filtro já ultrapassou a amostra inicial do trecho estacionário do sinal. Este transitório tem comprimento diretamente relacionado ao comprimento da resposta ao impulso do filtro.

De modo geral, os transitórios podem ser processados a ponto de não interferirem no problema de transcrição. Por exemplo, podemos cortar o transitório de subida através de um pré-processamento de detecção de inícios de notas, como observado nas Figs. 7 e 8, reduzindo-se, assim, a interferência entre notas subseqüentes numa mesma banda.

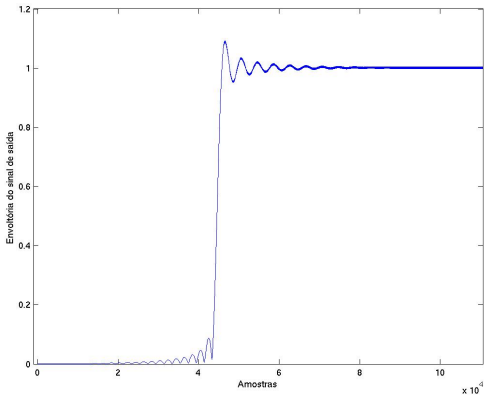

Figura 7: Exemplo 2 - Amplitude da resposta completa de uma banda do FRM-CMFB a uma senóide.

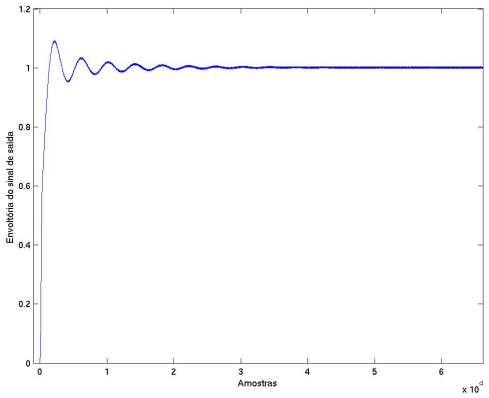

Figura 8: Exemplo 2 - Amplitude da resposta sem o transitório de subida de uma banda do FRM-CMFB a uma senóide.

Exemplo 3 - Freqüência Variável:

Uma forma simples de visualizar a acurácia de representação que se pode alcançar com os bancos de filtros é testá-los com uma senóide de freqüência variável. Neste exemplo, usamos uma senóide com sua freqüência variando linearmente de (f_c-20)Hz a (f_c+20)Hz, onde f_c é a freqüência central de uma banda do FRM-CMFB. Com isto, o sinal analisado consistia numa única freqüência variando no tempo ao longo de três bandas distintas do banco de filtros.

A parte superior da Fig. 9 mostra a variação de freqüência presente no sinal, atravessando as linhas horizontais que delimitam as bandas dos filtros do banco. A parte inferior da Fig. 9 mostra a envoltória das respostas a este sinal quando convoluído com os filtros de interesse neste exemplo. O resultado mostra como o banco de filtros foi capaz de perceber a banda correta da freqüência do sinal de entrada, inclusive determinando corretamente os momentos nos quais a freqüência mudava de banda dentro do banco de filtros.

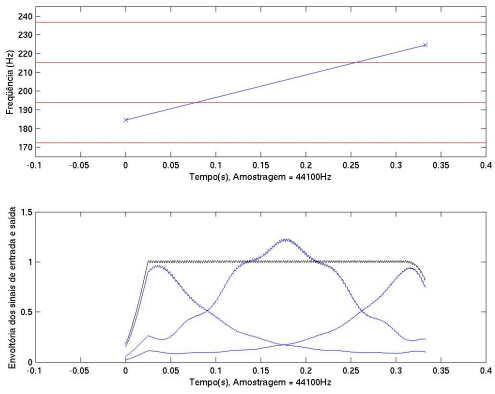

Figura 9: Exemplo 3 - (a) Variação da freqüência do sinal de entrada (as linhas horizontais representam limites de bandas no banco de filtros); (b) Envoltória de amplitude nas saídas das bandas correspondentes à parte (a).

VII - CONCLUSÃO

Neste trabalho, expusemos algumas premissas teóricas do problema da transcrição musical. Discutimos, ainda, o uso de bancos de filtros no reconhecimento de notas musicais, em particular a estrutura FRM-CMFB, ressaltando sua vantagem em relação à DFT quanto à seletividade no domínio da freqüência. Por fim, apresentamos exemplos práticos do problema da transcrição musical, enfatizando os aspectos da resolução no domínio da freqüência e dos transitórios no domínio do tempo.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] A. Klapuri, *Automatic Transcription of Music*, M.Sc. Thesis, Tampere University of Technology, Finland, Nov. 1997.

[2] S. R. Diniz, E. A. B. da Silva, S. L. Netto, *Digital Signal Processing: System Analysis and Design*, Cambridge, UK, 2002.

[3] P. S. R. Diniz, L. C. R. de Barcellos, S. L. Netto, "Design of cosine-modulated filter bank prototype filters using the frequency-response masking approach," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA, May 2001.

[4] S. L. Netto, P. S. R. Diniz, L. C. R. de Barcellos, "Efficient implementation for cosine-modulated filter banks using the frequency-response masking approach," *Proc. IEEE International Symposium on Circuits and Systems*, Scottsdale, AZ, USA, vol. III, pp. 229-231, May, 2002.

[5] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, USA, 3.ed.., 1991.

[6] T. F. Quatieri, R. J. McAulay, "Audio signal processing based on sinusoidal analysis/synthesis," in *Applications of Digital Signal Processing to Audio and Acoustics*, eds. M. Kahrs, K. Brandenburg, Kluwer, 1998.

[7] S. W. Foo, W. T. Lee, "Application of fast filter bank on transcription of polyphonic signals," to appear in *Journal on Circuits, Systems and Computers*, vol. 12, no. 5, Oct. 2003 (expected).