Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 337–341, April 2–5, 2001 - Rio de Janeiro - RJ - Brazil

Discriminação Neural de Instrumentos Musicais Baseada no Espectro

Cristiano Nogueira dos Santos, Luiz Pereira Calôba, Luiz Wagner Pereira Biscainho
Laboratório de Processamento de Sinais
DEL/EE & PEE/COPPE, Universidade Federal do Rio de Janeiro
C.P. 68564, Rio de Janeiro, RJ - Brazil
E-mails: {csantos, caloba, wagner}@lps.ufrj.br

Abstract

Some works have recently approached the problem of musical instrument recognition. This paper presents a first step towards a pitch independent recognition system of which the main target is to employ a reduced set of input parameters. The problem was constrained to instrument identification from emitted individual notes in steady state. A neural network receives as inputs the relative spectral magnitudes of the harmonic components of each sound sample, extracted by DFT. Training and testing data were obtained by recording synthesized (midi) notes for eight instruments from the reed family. Validation tests used audio samples from real recordings with success.

1. Introdução

As técnicas e tecnologias ligadas ao processamento de sinais têm passado por grandes avanços nos últimos anos. O estudo de sinais de áudio vem se beneficiando com isso e, em vários centros acadêmicos espalhados pelo mundo, as pesquisas se intensificam. O reconhecimento de voz, de fala e de objetos através dos sons produzidos por eles é um dos enfoques possíveis. Um dos principais objetivos desta área de pesquisa, hoje, é construir uma máquina que possa aprender a reconhecer e separar fontes sonoras em ambientes acusticamente complexos. Até o presente momento, nenhum resultado convincente foi apresentado com um grau de competência comparável ao do ouvido.

Desejando-se alcançar um sistema computacional que possa atuar como o aparelho auditivo humano e, futuramente, com qualidade ainda superior, as tentativas de se descobrir as características acústicas que nos habilitam a identificar sons de origens distintas têm sido intensas. Essas características formam grupos de parâmetros, que podem ser usados como entradas para uma rede neural responsável por reconhecer a fonte de excitação sonora.

Uma das vertentes desse campo de estudo visa à identificação de instrumentos musicais. Dentre os benefícios que desses trabalhos podem resultar estão o auxílio à separação de canais para remasterização e simulação de realidade virtual, além de auxiliar profissionais da música no estudo de obras musicais complexas. As publicações nessa área [1, 2, 3] vêm-se utilizando de conjuntos extensos de parâmetros, o que dificulta e prolonga o pré-processamento dos dados na entrada do sistema de discriminação.

Este trabalho propõe o uso de uma rede neural que aprenda a distinguir instrumentos musicais através de um único grupo de parâmetros dentre os diversos propostos na literatura, procurando analisar sua importância, dada a sua simplicidade e conseqüente alta velocidade de obtenção.

2. Timbre

As teorias divergem quanto à definição de quais sejam as peculiaridades que definem a chamada "cor" do som, indo das mais discretas com apenas um parâmetro até as mais detalhadas, com dezenas de especificações e suas diversas formas de análise. Em sua maioria, porém, trabalham com alguns aspectos mais relevantes. São eles a envoltória do sinal no tempo, assim como a do seu espectro de freqüência; o centróide do espectro, as características de ressonância, o assincronismo de tempo de ataque das componentes de alta e baixa freqüência, a modulação no tempo e em freqüência, além da freqüência fundamental em Hertz.

A análise de envoltória no tempo trata o acompanhamento da trajetória da amplitude do sinal. Ela traz a informação sobre a fonte de excitação e o corpo de ressonância. Alguns instrumentos têm um tempo de subida e decaimento próprio, dado seu grau de acoplamento entre o material vibrante e o corpo ressonante. A forma de se tocar o instrumento, claramente, também tem influência na forma de onda, e indica a família de instrumentos característica dos tempos de subida, descida e decaimento.

A modulação em amplitude surge em alguns tipos de instrumentos como "blips" não harmônicos, em baixa freqüência. Em outros casos, aparece periodicamente, sendo aí denominadas de "tremolo". Assim, podemos notar variações periódicas ou aleatórias, o que nos dá uma boa propriedade de discriminação. A modulação em freqüência apresenta os mesmos casos, sendo as aleatórias chamadas de "jitter" e as periódicas, de "vibrato". Esses tipos de variação podem variar muito em grau, de um instrumento para outro.

É sabido que o centróide do espectro tem alta correlação com as qualidades subjetivas de "brilho" do som produzido [1]. Sua variação no tempo também pode representar um bom fator de distinção. Outra observação de grande valor é a relação entre o centróide e a intensi-

dade, pois quanto mais altos são os tons, mais “brilho” eles adquirem, ou seja, o centróide se desloca para uma freqüência mais elevada.

Essas e ainda outras características envolvidas no estudo de timbre podem ser divididas entre o grupo das temporais e o das espectrais. Dentre os trabalhos conhecidos ligados ao reconhecimento de instrumentos musicais, há os que se utilizam de conjuntos com características temporais extraídas da envoltória de energia média quadrática de termo curto ou características temporais calculadas a partir das saídas de um correlograma log-lag [1] ou ainda coeficientes cepstrais. Eronen e Klapuri [3], por sua vez, lançaram mão tanto de características temporais quanto das espectrais.

Levando-se em consideração que o tempo de processamento inclui a prévia extração de dados e os cálculos para a montagem do conjunto de entradas da rede neural, e que da redução desse conjunto depende a eficiência de aplicações como a transcrição musical automática, propõe-se aqui um grupo de parâmetros mais simples. Partindo da idéia de que a simples sustentação de uma nota já é capaz de dizer que instrumento a está emitindo, tanto a músicos profissionais como a amantes da arte, intui-se que a principal fonte de informação está presente no espectro. Como o objetivo, aqui, é chegar à menor quantidade de dados possível, procuramos uma forma razoável de ter a envoltória espectral bem representada, buscando a verificação de sua eficiência.

Tendo em vista as últimas observações, decidiu-se realizar o reconhecimento baseado, exclusivamente, em uma análise da representação em freqüência do sinal, dada pela transformada de Fourier discreta. Este método demanda grande quantidade de amostras do sinal, para uma boa resolução espectral. Fazer com que a rede trate de muitos pontos não é favorável ao aumento da velocidade. Então, sofisticando um pouco o pré-processamento do sinal, podemos definir o conjunto de dados de entrada como sendo composto apenas dos principais elementos do espectro.

Tratando-se de sinais periódicos ou quase-periódicos, como são os musicais, podemos considerar que temos apenas informações relevantes em senóides, em forma de harmônicos; uma visualização de alguns instrumentos, nas Figuras de 1 a 4, ajuda a compreender a idéia. Encontrar uma relação entre as intensidades desses harmônicos e a de sua fundamental pode ser um bom conjunto de parâmetros para a discriminação.

3. Dados Utilizados

A representação em freqüência de um sinal emitido por um instrumento depende do tipo, do fabricante e do exemplar do instrumento, da ação do intérprete e das notas emitidas, entre outros parâmetros mais. Uma descrição razoavelmente completa exigiria que se analisasse nota por nota, buscando condições favoráveis de estacionaridade, tocadas pelo maior número possível de intérpretes, a partir de gravações em que houvesse um

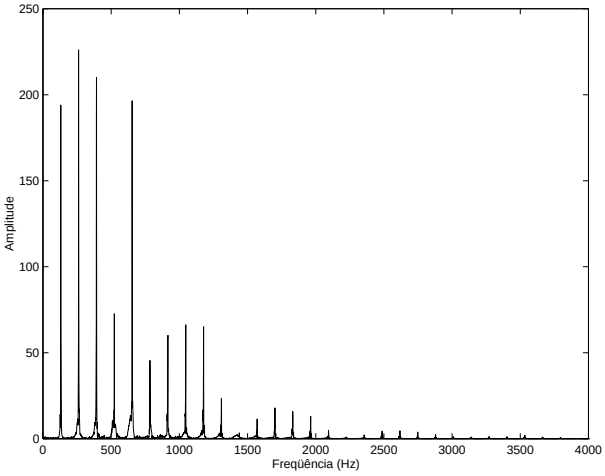

Figura 1: Espectro de uma nota de sax-soprano

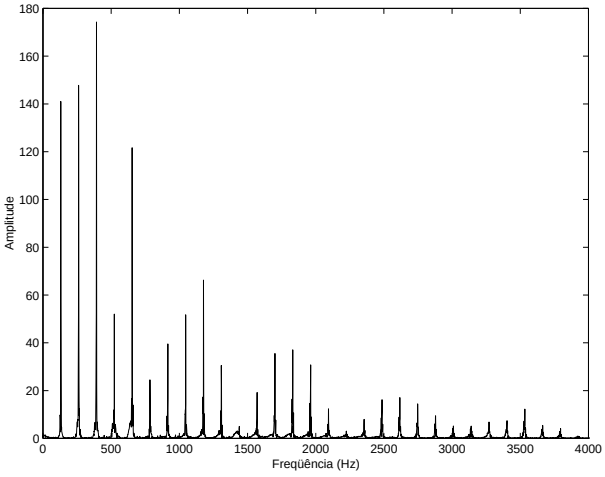

Figura 2: Espectro de uma nota de sax-alto

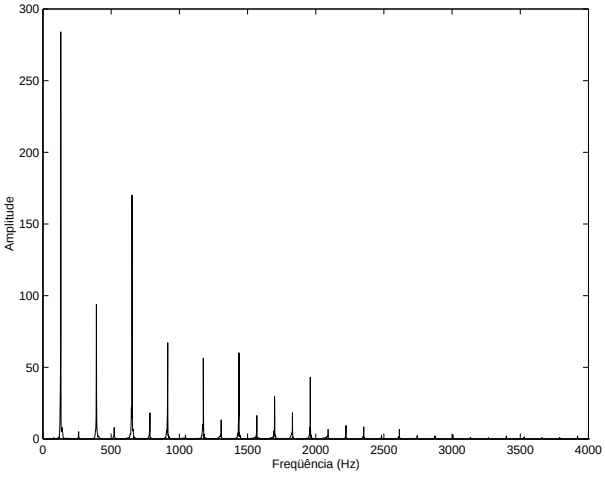

Figura 3: Espectro de uma nota de clarineta

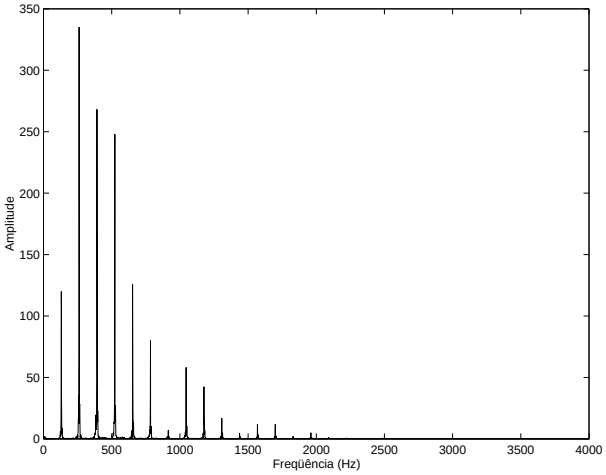

Figura 4: Espectro de uma nota de oboé

único instrumento a ser analisado, para se conseguir chegar às suas características.

A grande quantidade de recursos exigida impôs que se limitasse essa abrangência, ainda tentando garantir a generalidade do conjunto de dados. Nesse ponto, pareceu natural utilizar extratos de gravações de boa qualidade de discos compactos. Entretanto, as dificuldades envolvidas na obtenção de um conjunto representativo e condicionado de dados sugeriram outra simplificação. Em lugar de recortar trechos pré-gravados, optou-se por empregar amostras sintéticas de alta qualidade, em formato midi. A vantagem desta decisão foi permitir o uso de conjunto um pouco maior de instrumentos musicais e alturas de notas emitidas; o que se perdeu foi a variabilidade de formas de emissão. O enfoque foi, portanto, direcionado para a comparação dos espectros característicos típicos dos instrumentos sob análise.

A avaliação do desempenho do parâmetro escolhido para identificação direcionou a escolha do grupo de instrumentos a serem analisados. Assim, foram analisados apenas instrumentos de uma mesma família, a de sopros com palheta. Em número de oito, foram eles o sax-soprano, sax-alto, sax-tenor, sax-barítono, oboé, corne inglês, fagote e clarineta. As notas foram escolhidas dentro de um conjunto de trinta e seis, entre 130Hz e 1000Hz.

A aquisição de trechos de notas musicais, agora, consiste na geração das notas desejadas pelos instrumentos sintetizados, com pleno controle da duração (que garanta a estacionaridade) e da intensidade. O procedimento de obtenção consistiu em se definir as notas e suas durações em um software adequado ao som .mid, e gravá-las para o formato .wav, a partir de onde foram extraídos os vinte primeiros harmônicos, contando com o primeiro ou fundamental, de cada nota gravada. Em seguida foram feitas as normalizações das intensidades de cada harmônico em relação à da sua fundamental. Contudo, verificou-se que alguns instrumentos como os saxes, por exemplo, possuíam harmônicos com intensidade maior que a fundamental, assim a normalização passou a ser em relação ao harmônico de maior intensidade.

Para a validação do sistema, empregou-se um conjunto de sinais reais. Para tanto, foi necessário o recorte de trechos adequados de uma gravação. Tomando-se o cuidado de buscar notas isoladas.

4. Implementação

A arquitetura utilizada para a rede foi a *multilayer perceptron* [4, 5], com três camadas, sendo a intermediária testada com diferentes números de neurônios, demonstrando não serem necessários mais que 20 neurônios, para 20 harmônicos. Para 10 e 5 harmônicos, os testes mais eficazes foram aqueles em que se usou o número de neurônios da camada intermediária igual ao dobro do número de harmônicos. A camada de saída esteve sempre composta de oito neurônios, um para cada classe de instrumento musical. Todos os neurônios eram do tipo sigmóide com limites de 0 e 1. Os vetores de entrada variaram seu conjunto entre 20, 10 e 5 componentes, correspondendo aos respectivos primeiros harmônicos.

Inicialmente, o conjunto de dados de entrada se encontrava em ordem da nota mais grave para a mais aguda em cada instrumento e em ordem de instrumentos. Das 36 notas de cada um deles (3 oitavas, com 12 notas cada), um primeiro grupo formado pelas 18 primeiras notas foi utilizado para treinamento; o segundo grupo de 18 notas serviu para teste. Os resultados não foram muito bons, a visualização dos espectros mostrou que já que, para um mesmo instrumento, o "padrão" espectral (amplitudes relativas dos harmônicos) não é preservado ao longo das notas, alterando-se gradualmente com a escala cromática; daí a necessidade de se embaralhar as notas para que não haja especialização da rede no primeiro grupo, que apresenta estatísticas ligeiramente diferentes do segundo.

O método de retropropagação de erro foi usado para o treinamento da rede, por batelada. Foram testadas taxas de aprendizado de 0,005, 0,01, 0,02 e 0,04, apresentando melhor desempenho a taxa fixa de 0,02. Cada um dos melhores casos de cada quantidade de harmônicos foi examinado até onde haveria treinamento em excesso, testando-se a saída da rede a cada 100 épocas de treinamento, e verificando-se a porcentagem de acerto.

Os resultados obtidos utilizando os 20 primeiros harmônicos corresponderam às expectativas, como se observa na Tabela 1, que deve ser entendida da seguinte forma: na primeira coluna figuram os instrumentos musicais sob análise, cada um dos quais associado a um neurônio da camada de saída; na primeira linha, estão os instrumentos identificados após o teste. Desconsiderando inicialmente os saxes, pode-se dizer que os instrumentos foram bem reconhecidos. Do mesmo modo, o desempenho para o sax-soprano e o sax-alto foi bom. Um comportamento curioso se deu com o sax-tenor e o sax-barítono: ambos obtiveram índice de acerto inferior e, com freqüência, se confundem mutuamente. A posterior observação de seus espectros de freqüência de diferentes notas mostrou que, por uma peculiaridade do sistema de síntese escolhido para geração das amostras,

várias delas continham distribuição harmônica idêntica, sendo que, coincidentemente, as que foram treinadas para um instrumento eram idênticas às que seriam testadas como do outro. Portanto, para o cálculo do índice de acerto médio, que foi de 94,44%, foram desconsiderados os resultados desses dois instrumentos.

A Tabela 2 faz referência aos testes com apenas os 10 primeiros harmônicos de cada nota. Os índices de acerto, praticamente iguais aos dos testes com 20 harmônicos, levam a crer que é suficiente trabalhar com este número reduzido de componentes do sinal de áudio para essa família de instrumentos. A confusão entre o sax-barítono e o sax-tenor ocorreu novamente aqui, pelas mesmas razões descritas no parágrafo anterior. O índice de acerto médio foi de 93,52%.

A Tabela 3, por fim, apresenta os resultados dos testes com os 5 primeiros harmônicos. Aqui, já aparecem as dificuldades na identificação de alguns instrumentos. Para metade dos casos existe uma queda no desempenho da rede neural, particularmente no caso do sax-alto, cujo índice de acerto passou a beirar os 50%. O índice de acerto médio desceu a 85,19%.

5. Validação

Como exemplo de teste com sinal real, a partir de uma gravação de sax-alto tocando sem acompanhamento, foram recortadas 19 notas que apresentassem condições favoráveis à análise espectral, ou seja, de duração não tão curta que comprometesse a resolução espectral, nem tão longa que comprometesse a estacionaridade. Os testes partiram do processamento baseado em 10 harmônicos. Os respectivos resultados são apresentados na Tabela 4. Os números mostram um baixo grau de acerto, havendo confusão até mesmo com a clarineta. Nesse ponto, decidiu-se retornar ao processamento baseado em 20 harmônicos, cujos resultados aparecem na Tabela 5. Aqui, quase não houve erro na identificação do sax-alto. Provavelmente, para uma boa discriminação de saxes, como mostram os espectros das Figuras 1 e 2, é relevante a presença de mais harmônicos, devido aos seus harmônicos de ordem mais baixa serem de configuração semelhante.

6. Conclusões

Este trabalho propôs e avaliou um método de reconhecimento de instrumentos musicais baseado em rede neural. Tendo como principal objetivo fugir da multiplicidade de parâmetros utilizados na descrição dos sons, típica dos sistemas usuais, o sistema aqui proposto emprega apenas as informações espectrais referentes a cada nota emitida, em estado permanente.

O treinamento e os testes da rede foram realizados a partir de notas sintetizadas de instrumentos de sopro da família das palhetas, confirmando a importância das amplitudes relativas dos harmônicos no reconhecimento dos instrumentos. Trechos retirados de gravação real foram utilizados para validar o sistema, com sucesso.

Essa forma de processamento pode fazer parte de um sistema mais completo, capaz de separar, reconhecer e descrever linhas musicais múltiplas. Sua continuação natural envolve aprimorar os métodos de extração das informações, avaliar melhor a arquitetura da rede, enriquecer os testes com mais sinais extraídos de gravações reais e implementar gradualmente o sistema completo de reconhecimento.

Aplicações potenciais, além daquelas diretamente musicais, incluem a obtenção artificial de diversos canais a partir de matrizes mono para remasterização binaural ou ambientação espacial (“surround”).

Agradecimentos

Nossos agradecimentos a Fábio Pacheco Freeland (Universidade Federal do Rio de Janeiro, PEE-COPPE), Marcos da Rocha Vassali (Universidade Federal do Rio de Janeiro, PEE-COPPE) e Paulo Antonio Andrade Esquef (Helsinki University of Technology, Lab. of Acoustics and Audio Signal Proc.) pela sua inestimável colaboração neste trabalho.

Referências

[1] K. D. Martin. Toward automatic sound source recognition: Identifying musical instruments. In *1998 NATO Advanced Study Institute on Computational Hearing*, Il Ciocco, Italy, July 1998. Advanced Study Institute on Computational Hearing.

[2] K. D. Martin. Musical instrument identification: A pattern-recognition approach. In *136th Meeting of the Acoustical Society of America*, Norfolk, VA, USA, Oct. 1998. Acoustical Society of America.

[3] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Proceedings of the IEEE 2000 International Conference on Audio Speech and Signal Processing*, Istambul, Turkey, June 2000. IEEE.

[4] P. D. Wasserman. *Neural Computing: Theory and Practice*. Van Nostrand Reinhold, 1989.

[5] S. Haykin. *Neural Networks*. McGraw-Hill, 2nd edition, 1998.

Tabela 1: Discriminação para 20 harmônicos

	sax-soprano	sax-alto	sax-tenor	sax-barítono	oboé	corne inglês	fagote	clarineta
sax-soprano	15/18	-	-	-	1/18	-	-	2/18
sax-alto	-	17/18	-	1/18	-	-	-	-
sax-tenor	-	2/18	4/18	12/18	-	-	-	-
sax-barítono	-	3/18	6/18	9/18	-	-	-	-
oboé	-	-	-	-	18/18	-	-	-
corne inglês	-	-	-	-	-	18/18	-	-
fagote	-	-	-	-	1/18	1/18	16/18	-
clarineta	-	-	-	-	-	-	-	18/18

Tabela 2: Discriminação para 10 harmônicos

	sax-soprano	sax-alto	sax-tenor	sax-barítono	oboé	corne inglês	fagote	clarineta
sax-soprano	15/18	1/18	-	2/18	-	-	-	-
sax-alto	-	15/18	-	3/18	-	-	-	-
sax-tenor	-	4/18	7/18	7/18	-	-	-	-
sax-barítono	-	4/18	10/18	4/18	-	-	-	-
oboé	-	-	-	-	18/18	-	-	-
corne inglês	-	-	-	-	-	18/18	-	-
fagote	-	-	-	-	-	1/18	17/18	-
clarineta	-	-	-	-	-	-	-	18/18

Tabela 3: Discriminação para 5 harmônicos

	sax-soprano	sax-alto	sax-tenor	sax-barítono	oboé	corne inglês	fagote	clarineta
sax-soprano	16/18	-	-	-	-	-	-	2/18
sax-alto	2/18	10/18	-	1/18	1/18	2/18	-	2/18
sax-tenor	2/18	3/18	2/18	11/18	-	-	-	-
sax-barítono	2/18	5/18	3/18	8/18	-	-	-	-
oboé	-	-	1/18	-	15/18	-	2/18	-
corne inglês	-	-	-	-	-	18/18	-	-
fagote	-	-	-	-	-	1/18	17/18	-
clarineta	-	-	-	-	-	-	-	18/18

Tabela 4: Discriminação com 10 harmônicos para sinais reais de sax-alto

	sax-soprano	sax-alto	sax-tenor	sax-barítono	oboé	corne inglês	fagote	clarineta
sax-alto	-	10/19	5/19	1/19	-	-	-	3/19

Tabela 5: Discriminação com 20 harmônicos para sinais reais de sax-alto

	sax-soprano	sax-alto	sax-tenor	sax-barítono	oboé	corne inglês	fagote	clarineta
sax-alto	-	17/19	2/19	-	-	-	-	-