Aplicação de MFCC para modelar sons de instrumentos musicais

Alan R. Fachini
Center of Technological Sciences (CCT)
Santa Catarina State University (UDESC)
Joinville, SC, Brazil
Alan.fachini@joinville.udesc.br

Milton Roberto Heinen
Computer Engineer Dept.
Federal University of Pampa (UNIPAMPA)
Bagé, RS, Brazil
Milton.heinen@unipampa.edu.br

***Resumo*—Devido à grande quantidade de conteúdo multimídia atualmente disponível, realizar a busca em uma base de dados multimídia pode se tornar uma tarefa complicada, visto que a abordagem tradicional presente na maioria dos sistemas consiste em recuperar documentos através de determinadas palavras-chave recuperando informações através do nome do arquivo ou de metadados. Este artigo apresenta o uso do vetor de características MFCC em conjunto com as técnicas de aprendizado de máquina Perceptrons de múltiplas camadas (MLP) e mapas auto-organizáveis (SOM) para categorizar sons de instrumentos musicais.**

Palavras chave—redes neurais; mapas auto-organizáveis; processamento de audio; reconhecimento de padrões.

I. INTRODUÇÃO

Nos últimos anos, houve uma crescente atividade de pesquisas voltadas para a análise de sinais baseada em conteúdo de áudio. A necessidade de interagir com grandes bibliotecas de áudio é uma demanda em várias áreas (FOOTE, 1999). Produtores de filmes, animação e conteúdo para a televisão necessitam interagir com uma vasta quantidade de efeitos de som presentes em bibliotecas multimídia. O mesmo se aplica á criação de um jogo, que possui centenas de sons diferentes utilizados. Compositores e DJs, que trabalham com a produção de música digital utilizam grandes coleções de sons para criar suas músicas (TZANETAKIS, 2002). Esta aplicação também é interessante para a busca por uma música em uma estação de rádio, busca de vídeos por similaridade de áudio na biblioteca de uma emissora de televisão, além de outras aplicações voltadas para a web e a imensa quantidade de informações atualmente disponível.

Atualmente a grande maioria dos sistemas de busca de áudio disponível indexa os arquivos de áudio através de informações como nome do arquivo ou metadados. Porém, nos últimos anos, vem crescendo o interesse no desenvolvimento de sistemas de classificação automática de sinais de áudio através da análise e extração de características e, alguns sistemas e técnicas foram propostos.

II. ESTADO DA ARTE

Diferentes descritores de áudio vêm sendo estudados e utilizados como vetores de características. Alguns dos mais citados, por exemplo, são os descritores de características psicoacústicas como o timbre, envelope temporal e espectral e Coeficientes Mel-Cepstrais (MFCCs) (WOLD et al., 1996; TZANETAKIS; COOK, 2000; LOGAN, 2000; BRENT, 2010; HERRERA et al., 2002).

Feiten e Günzel (1994) apresentam como proposta utilizar uma Rede Neural Auto-Organizada para organizar sons. Esse mapa pode ser utilizado para quantizar a representação vetorial dos sons e ser utilizado para identificar classes no espaço de sons. Sons similares são mapeados em vizinhanças, e sons muito diferentes são separados por largas distâncias. A extração de características é realizada e as principais frequências audíveis pelo ouvido humano são extraídas pelo filtro de frequência nark, o qual é utilizado como vetor de entrada da rede neural. Ao final do artigo, os autores concluem que a proposta mostrou-se eficiente na indexação dos sons, e propõem estudos futuros relacionados com características psicoacústicas como o timbre.

Wold et al. (1996) propõe um sistema que recebe como entrada um arquivo de áudio, que tem seu conteúdo comparado com os arquivos presentes na base de dados. Os autores descrevem as características do sistema Muscle Fish, o qual analisa a similaridade entre dois sinais de áudio utilizando características acústicas: volume, ruído, frequência fundamental, brilho (brightness), largura de banda e harmonicidade. O vetor de características é composto pela média, variância e auto-correlação para cada aspecto do som analisado. Para a classificação dos sons é utilizada a distância euclidiana. A distância é comparada com um limiar para determinar se o som está ou não na classe, sendo colocado na classe onde a distância for menor.

Foote (1997) descreve um sistema de recuperação de arquivos de áudio utilizando similaridade acústica. O sistema extrai um vetor de características contendo o *Mel-Frequency Cepstral Coeficients* (MFCCs), construindo uma árvore de decisão quantizada (*tree-based quantizer*). Esta operação requer supervisão para que os dados de treinamento sejam rotulados. A árvore automaticamente particiona o espaço de características em regiões que possuem diferentes classes de população. Quando um áudio é utilizado para realizar a busca por similaridade, um *template* é gerado e comparado com a coleção de *templates* existentes, o que retorna a similaridade com cada arquivo de áudio na coleção. O algoritmo utiliza a

distância euclidiana. Dessa forma, o resultado ordenado por similaridade é apresentado.

Herrera et al. (2002) apresentam uma avaliação comparativa para a classificação automática de um banco de dados contendo sons de bateria.É proposto o uso de um vetor de vinte características, com descritores de envelope temporal e espectral, e diferentes técnicas como classificação baseada no vizinho mais próximo (K-Nearest Neighbors) e Árvore de Decisão são testadas. Os autores relatam que em seus testes conseguiram 99% de acerto.

III. *Mel Frequency Cepstral Coefficients*

Mel-Frequency Cepstral Coefficients é tipicamente utilizado para tarefas de reconhecimento de fala, mas vem sendo utilizado em aplicações musicais (TZANETAKIS; COOK, 2002).

O MFCC leva em conta a percepção não linear do som pelo ouvido humano. O que torna o uso de MFCC interessante é o fato de sua aplicação reduzir um espectro de 1024 pontos para cerca de 15 a 40 pontos que podem ser utilizados para verificar a similaridade ou distinção de sons (BRENT, 2010).

O processamento do Mel-Frequency Cepstral Coefficients é descrito pelo diagrama da Figura 1. Inicialmente é realizado o processo de janelamento e aplicada a DFT. A amplitude da Transformada de Fourier é filtrada por janelas triangulares na escala Mel e então aplica-se o logaritmo. A Transformada Discreta de Cosseno é aplicada e os Coeficientes Mel-Cepstrais são as amplitudes resultantes (LOGAN, 2000).

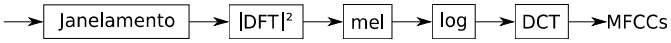

Fig. 1. Mel-Frequency Cepstral Coefficients

O primeiro passo para realizar a computação dos Coeficientes é segmentar o sinal em quadros de igual tamanho e multiplica-se os quadros por uma função de janelamento *w(n)* onde *n* é um quadro qualquer do sinal. A janela de Hamming é indicada para tarefas de processamento de sinais de áudio (Equação1):

$$w(n) = 0.54 - 0.46\ \cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \leq n \leq N-1 \tag{1}$$

onde *N* é o tamanho da janela. A Figura 2 mostra a forma de onda da janela.

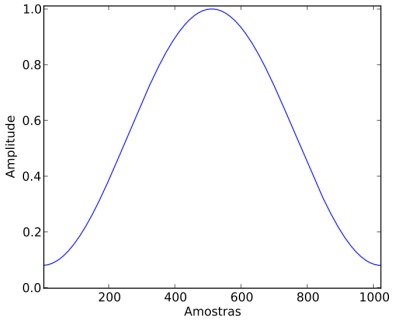

Fig. 2. Janela de Hamming

O janelamento do sinal é dado pela Equação 2:

$$x_j(n) = x(n)w(n) \tag{2}$$

onde $x_j(n)$é o sinal de saída, janelado, *x(n)* o sinal de entrada e *w(n)* a janela a ser aplicada.

Após o janelamento do sinal, a Transformada Discreta de Fourier (DFT) é processada. A DFT (Equação 3) recebe como entrada um sinal de tamanho *N*, produzindo, como saída, um sinal de mesmo tamanho representando os coeficientes espectrais, ou espectro.Nesta equação *f(k)*é o sinal de entrada, $e^{-j2\pi}$é a função base que define os valores complexos para cada ponto *F(n)* no domínio de frequência.

$$F(n) = \sum_{k=0}^{N-1} f(k)e^{-j2\pi\frac{kn}{N}} \tag{3}$$

A escala Mel é uma escala psicoacústica que explora a relação de percepção da frequência fundamental entre dois tons, criada a partir do estudo da dinâmica do sistema auditivo humano. A unidade de medida Mel (em referência a melodia) refere-se a frequência subjetiva de tons puros percebida pelo ouvido humano (BRENT, 2010).O mapeamento entre a frequência f_{Hz} e a frequência percebida f_{mel} é dada pela Equação 4 e gera a curva apresentada na Figura 3:

$$f_{mel}(f_{Hz}) = 1127.01048\ \log\left(1 + \frac{f_{Hz}}{700}\right) \tag{4}$$

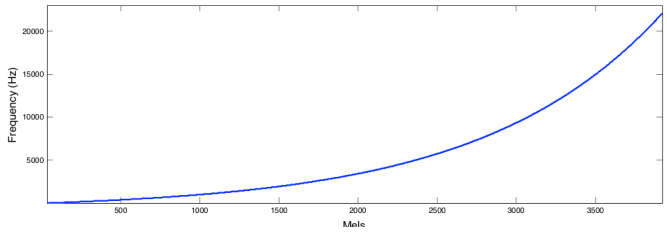

Fig. 3. Escala Mel (BRENT, 2010)

Para criar o banco de frequências Mel, utilizam-se filtros passa-banda com envelopes triangulares, com o centro dos filtros espaçados de acordo com a escala Mel, enfatizando as frequências próximas ao centro do filtro (onde ocorre a mudança perceptiva de tom) e atenuando as frequências vizinhas. A Figura 4 mostra a aplicação dos bancos de filtros.

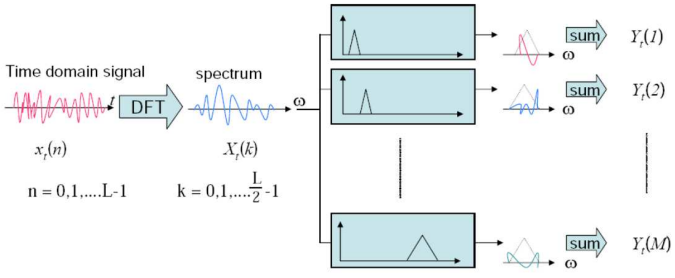

Fig. 4. Aplicação do banco de filtros (JURAFSKY, 2009)

O último passo para a obtenção dos Coeficientes Mel-Cepstrais é aplicar a Transformada Discreta do Cosseno (DCT) às componentes geradas pela aplicação do banco de filtros. Esta transformada difere da DFT pelo fato de ser aplicada somente a sequências reais. A DCT-2 é definida pela Equação5:

$$DCT_2(n) = \sum_{k=0}^{N-1} f(k)\ \cos\left(\frac{\pi}{N}(k+\frac{1}{2})n\right) \quad (5)$$

IV. SISTEMA PROPOSTO

A Figura 5 define a arquitetura do sistema proposto, que é dividida em 4 partes principais: criar um banco de sons de instrumentos musicais, extração de características, sistema de classificação e sons classificados.

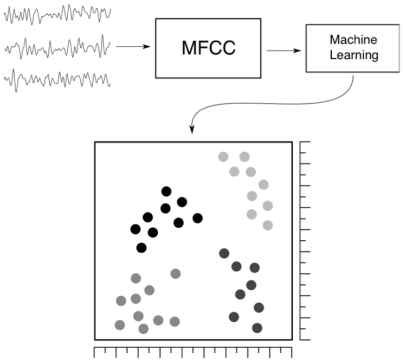

Fig. 5. Sistema Proposto

Para a implementação do sistema proposto escolheu-se utilizar o sistema de *plugins* Vamp. Este sistema foi escolhido por integrar em um único ambiente a visualização de dados através da aplicação Sonic Visualizer e extração de características através do programa Sonic Annotator (MARY, 2007), ficando a nosso cargo apenas escrever os scripts de extração de características. Os *plugins* são executados por um programa hospedeiro, recebendo sinais de áudio como entrada, realizando a extração de características e disponibilizando como saída estruturas de dados contendo informações sobre o sinal processado.

Para classificação dos dados foram utilizadas as técnicas de Rede Neural do tipo MLP (Multilayer Perceptron) e Self-Organizing Maps (SOM). Rede Neural do tipo MLP se enquadra na categoria de algoritmos de Aprendizado de Máquina Supervisionado e consiste em várias camadas de nós (neurônios) onde cada camada é conectada na próxima seguindo o padrão de um grafo direcionado. O SOM trata-se de um modelo de rede neural artificial baseado na descoberta de que a informação prevalecente em dados de entrada n-dimensionais pode ser mapeada em uma camada de neurônios uni ou bidimensional (Frigo 2010). O SOM consiste basicamente de duas camadas de neurônios: uma camada de entrada e uma camada de saída (camada de Kohonen). As entradas correspondem a vetores no espaço n-dimensional. Cada neurônio da camada de Kohonen apresenta se totalmente conectado com todas as entradas do mapa, possuindo um vetor de pesos sinápticos associado, também no espaço n-dimensional (Frigo apud Gonçalves 2009).

Para a realização dos testes, sons dos seguintes instrumentos foram utilizados: Cuíca (5), Triângulo e Sino (37), Bongo (57), Castanhola (20), Palmas (11), Chimbau (36), Hi-hat (18), Caixa clara (48), Baqueta (7), Pandeiro (8), Tom-tom (32).Ao todo, 279 amostras foram utilizados para os testes, coletados das bases abertas de áudio Freesound (PROJECT, 2010a) e OLPC Sample Library (PROJECT, 2010b).

V. TESTES REALIZADOS

Entre os instrumentos musicais percussivos escolhidos, identificam-se instrumentos que possuem timbres diferentes e outros parecidos. Mesmo dentro de um conjunto de sons de um determinado instrumento, podemos verificar uma grande variação de timbres. Isso deve-se principalmente ao fato de a maioria dos instrumentos permitir diferentes afinações e configurações ou mesmo possuírem derivações construídas com diferentes materiais.A Figura 6 mostra gráficos dos primeiros 40 coeficientes extraídos para alguns instrumentos.

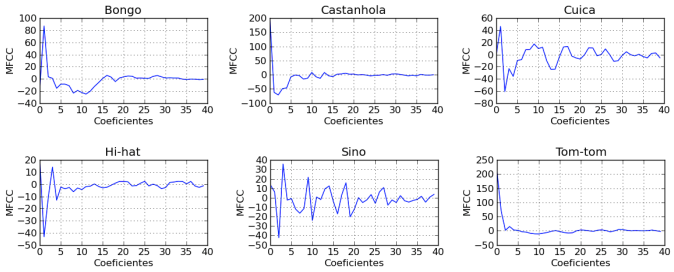

Fig. 6. MFCC dos intrumentos Bongo, Castanhola, Hi-hat, Cuíca, Tom-tom e Sino

A partir de uma análise dos gráficos, pode-se perceber os primeiros coeficientes são os mais discriminantes, isso deve-se ao fato de concentrarem os coeficientes das faixas de frequência com maior potência.

Para a implementação da Rede Neural MLP foi utilizada a biblioteca PyBrain com as configurações: 13 camadas de entrada, 19 camadas interna e 11 camadas de saída, com uma taxa de aprendizado de 0.1 e 70 épocas de treinamento. A quantidade de neurônios na camada interna foi definida a partir de testes realizadas para diversas configurações, onde 19 neurônios nessa camada obteve o melhor resultado de classificação. O Erro para o pior dos casos de teste foi de 16%. A Tabela 1 mostra a Matriz de confusão.

TABLE I. MATRIZ DE CONFUSÃO PARA OS TESTES UTILIZANDO MLP

#	Bongo	Cuíca	Sino	Castanholas	Palmas	Chimbau	Hi-hat	Baqueta	Caixa clara	Pandeiro	Tom-tom
Bongo	50	0	0	7	0	0	0	0	0	0	0
Cuíca	0	4	1	0	0	0	0	0	0	0	0
Sino	0	0	37	0	0	0	0	0	0	0	0
Castanholas	0	0	1	19	0	0	0	0	0	0	0
Palmas	0	0	0	5	6	0	0	0	0	0	0
Chimbau	0	0	0	2	0	34	0	0	0	0	0
Hi-hat	0	0	0	2	0	2	14	0	0	0	0
Baqueta	1	0	1	5	0	0	0	0	0	0	0
Caixa clara	3	0	0	12	0	0	0	0	31	0	2
Pandeiro	0	0	0	0	0	0	0	0	0	8	0
Tom-tom	0	0	0	5	0	0	0	0	3	0	23

Para os testes realizados com SOM, as configurações da matriz da rede ficou com dimenções 100x100, com learningrate de 0.05 e 400 épocas de treinamento. Para o pré-processamento dos dados de entrada da rede foi aplicado o técnica PCA (Principal Component Analysis) para reduzir a dimensão dos vetores de características extraídos das janelas dos sons de 13 para 3 dimensões e então concatenadas para serem utilizados como entrada da rede. A Figura 6 mostra a saída do SOM para uma das execuções de teste.

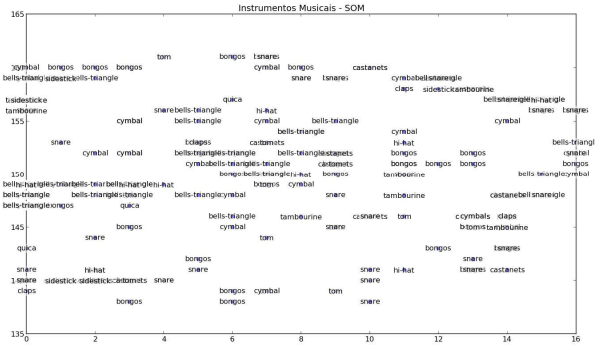

Fig. 7. Figura 6: Saída do SOM

A qualidade da classificação dos dados utilizando SOM não foi satisfatória e maiores detalhes sobre os resultados não foram levantados.

VI. CONCLUSÕES

Neste trabalho foi realizado o estudo o vetor de características MFCC para extração de característica de áudio a fim de caracterizar o timbre de instrumentos musicais.

O timbre é uma característica psicoacústica utilizada para descrever sons, percebido através de um processo cognitivo muito complexo que acontece através da interpretação dos sons pelo cérebro humano, não existindo uma forma simples de representá-lo. O vetor de características MFCC tem mostrado bons resultados em tarefas de classificação de sinais de áudio. Por não ser um descritor simples, encontra-se dificuldade em correlacionar os dados extraídos, porém podemos notar na Figura 6 que para diferentes instrumentos musicais temos diferenças nos vetores de características. A literatura também relata bons resultados de classificação utilizando este vetor de característica.

A classificação de sinais de áudio mostrou-se uma tarefa desafiadora. Após realizar os estudos sobre extração de características de áudio, constatou-se que os modelos estudados são capazes de descrever o som de um instrumento musical. O escopo do conjunto de testes selecionado para o desenvolvimento do sistema ficou restrito a sons de instrumentos percussivos, a fim de validar a proposta com um conjunto de testes reduzido, podendo ser estendido a outros tipos de sons em trabalhos futuros.

Foram obtidos bons resultados utilizando o algoritmo MLP, com uma taxa de erro de 16% para o pior caso. Porém, ao aplicar a algoritmo SOM os resultados não foram os esperados. Considera-se que o desafio para melhorar os resultados de classificação está na fase de extração e pós-processamento dos vetores de características, de forma a deixá-los mais expressivos e representarem melhor a característica estraídas dos conjuntos de testes. Sugere-se realizer trabalhos futuros nesse sentido e em testes utilizando outras técnicas de aprendizado de máquina.

REFERÊNCIAS

[1] BRENT, W. “Physical and Perceptual Aspects of Percussive Timbre”. Tese (Doutorado) — University of California, 2010.

[2] FEITEN, B.; GÜNZEL, S. “Automatic indexing of a sound database using self-organizing neural nets”. Computer Music Journal, MIT Press, v. 18, n. 3, p. 53–65, 1994.

[3] FOOTE, J. “An overview of audio information retrieval”. Multimedia Systems, Springer Berlin / Heidelberg, v. 7, p. 2–10, 1999. ISSN 0942-4962. 10.1007/s005300050106. Disponível em: <http://dx.doi.org/10.1007/s005300050106>.

[4] FOOTE, J. T. “Content-based retrieval of music and audio”. In: KUO, C.-C. J.; CHANG, S.-F.; GUDIVADA, V. N. (Ed.). [S.l.]: SPIE, 1997. v. 3229, n. 1, p. 138–147.

[5] FRIGO, O. “Classificação automática de imagens baseada em mapas auto-organizáveis”. TCC (Graduação) – UDESC, 2010.

[6] HAYKIN, S. “Redes Neurais: Princípios e prática”. 2. ed. Porto Alegre: Bookman, 2001.

[7] HERRERA, P.; YETERIAN, A.; GOUYON, F. “Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques”. Music and Artificial Intelligence, Springer, p. 69–80, 2002.

[8] LOGAN, B. “Mel frequency cepstral coefficients for music modeling”. In: BYRD, D. (Ed.). International Symposium on Music Information Retrieval (ISMIR) Proceedings. Plymouth, Massachusetts, USA: [s.n.], 2000. v. 28.

[9] MARY, C. for D. M. Q. “The Vamp audio analysis plugin system”. 2007. Disponível em: <http://vamp-plugins.org>. Acesso em: 03/11/2010.

[10] PROJECT, F. "Freesound. 2010. Disponível em: <www.freesound.org/>. Acesso em: 18/05/2011.

[11] PROJECT, O. L. P. C. “The Open Path Music Custom Sample Library for OLPC”. 2010. Disponível em: <wiki.laptop.org/go/Free sound samples>. Acesso em: 18/05/2011.

[12] TZANETAKIS, G.; COOK, P. “Musical genre classification of audio signals”. Speech and Audio Processing, IEEE Transactions on, v. 10, n. 5, p. 293 – 302, July 2002.

[13] WOLD, E.; BLUM, T.; KEISLAR, D.; WHEATEN, J. “Content-based classification, search, and retrieval of audio”. Multimedia, IEEE, IEEE, v. 3, n. 3, p. 27–36, 1996.