Sistema de reconhecimento de instrumentos musicais em reprodução simultânea

Tiago Araújo, aluno da Universidade de Aveiro

tiago.araujo@tvtel.pt

1-Introdução

O sistema apresentado visa o reconhecimento automático de instrumentos musicais e tem dois modos de operação: "offline" o resultado da classificação é apresentado no final do processamento de todo o ficheiro áudio; em tempo real onde a reprodução de som e a sua classificação são realizadas em simultâneo.

Este sistema é composto essencialmente por quatro blocos de processamento: pré-processamento, extracção de características, selecção de características e classificação.

O diagrama de blocos da figura1, representa o funcionamento do sistema de reconhecimento.

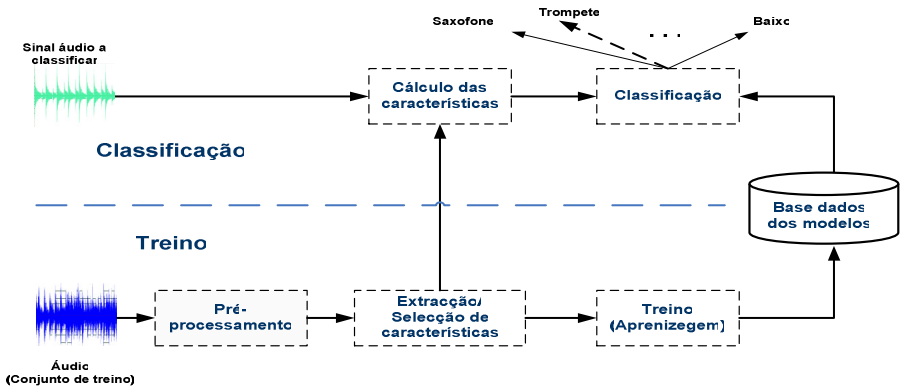


Figura 1: Arquitectura do sistema de reconhecimento de instrumentos musicais.

*2-*Caracterização do sistema

Nesta secção pretende-se explanar os conceitos mais relevantes envolvidos em cada um dos blocos.

2.1. Pré-processamento

Tem por objectivo "preparar" o sinal, para o bloco de extracção de características. O sinal à entrada é segmentado em "frames" para permitir amenizar dois problemas básicos; o problema do modo de funcionamento em tempo real e o facto de o som produzido pelos instrumentos não ser periódico (o processamento em segmentos permite ter uma percepção mais "localizada" das convoluções do sinal).

Para esta aplicação, optou-se por cortar o sinal em segmentos com 23,2 ms de duração (512 amostras @ 22050Hz ou 1024 amostras @ 44100Hz) com um salto de 256 ou 512 (sobreposição de 50%) conforme a frequência de amostragem (22050Hz ou 44100Hz).

2.2. Extracção de características

Pretende-se que a análise do sinal acústico conduza a uma representação que preserve toda a informação necessária para o processo de classificação, mas seja insensível a informação não relevante, favorecendo também uma acentuada redução da dimensão do espaço de representação acústica.

Nesta secção são definidos os métodos de cálculo e de representação numérica do sinal que preserve toda a informação necessária para o processo de classificação. O desafio passa por encontrar um vector de características de um dado instrumento reveladoras do seu timbre, favorecendo também uma acentuada redução da dimensão do espaço de representação acústica.

Foram implementados *três* grandes grupos de características: *psico-acústicas*, que simulam o comportamento do ouvido humano na discriminação de fenómenos acústicos; *características probabilísticas* que descrevem dependências entre acontecimentos sem relação aparente (recorrendo à análise estatística) e as características temporais que definem acontecimentos/eventos (ex. "onsets"), no domínio temporal.

2.2.1 Características psico-acústicas

Nesta subcategoria foram implementadas as seguintes características: *coeficientes MelCepstrais*; e*nergia das bandas* definidas segundo a escala *Bark* [2]; cálculo da *energia localizada* em cada banda *ERB* [1][2]; energia em *conjuntos de bandas adjacente*s (descriminação mais grosseira do modo como o espectro se comporta ao longo da frequência); *medida da variação da energia*, de cada banda, ao longo do tempo [3].

2.2.2 Características probabilísticas

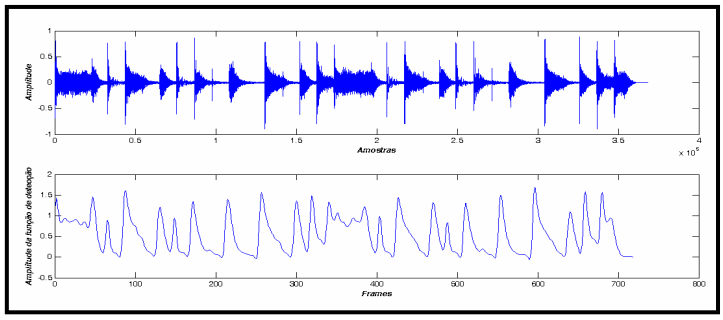
As características definidas foram: *média e variância das características* descritas acima, ao longo de um número predefinido de segmentos ("*frames*"); *Skewness* e *Kurtosis* (caracterizam respectivamente a assimetria da função de distribuição de probabilidade "*fdp*" e a sua largura quando comparada com a distribuição normal Gaussiana).

2.2.3 Características temporais

Nós, humanos conseguimos detectar e identificar instrumentos apenas pelo seu ataque, principalmente em misturas complexas.

Porém, a análise dos ataques de instrumentos acústicos é algo não trivial , dado que, normalmente, a região do sinal é não estacionária, onde as técnicas de análise "ditas" tradicionais (LPC, MFCC's, etc) dão maus resultados.

Assumindo secções de sons monofónicos (linhas melódicas, solos), primeiro é necessário saber quando começa a nota para a seguir analisar. Sem a detecção do início das notas, a análise pode conter uma parte da representação desfocada, com ataques, transitórios e regimes estacionários à mistura. Perante este problema surgiu a motivação para a implementação de um detector de onset's, como o objectivo de logo de seguida ser possível a determinação do "*atack-time"* dos instrumentos.



Para além da localização temporal dos "*onsets*", usa-se a função de detecção para descriminar instrumentos com base no ritmo ("*beat*"), e no nível de percussividade (muito percussivo => picos da função abruptos).

Figura 2: Exemplificação da transformação que a função de detecção realiza ao sinal.

Após o cálculo destas características, é necessário reduzir o número de vectores característica (um por cada "*frame*"). Esta necessidade deve-se à elevada carga computacional requerida pelos processos de treino/classificação (*secção 2.4*).

Para esta redução, agrupam-se todos os vectores característica em conjunto de 10 e acha-se a mediana de cada uma das características, nesse grupo. Cada um desses novos vectores representa um "*chunk*".

Com este processo, para além da redução do seu número (somatório do número de "*frames*" a dividir por 10), "suaviza-se" também os resultados (a análise "*frame*" a "*frame*" é demasiado especifica e por vezes, não contém informação suficiente para caracterizar o comportamento de um instrumento).

2.3. Selecção de características

Selecção de características ("*feature selection*") é o processo que determina um conjunto mais reduzido de características, melhorando a precisão da classificação, e por outro lado permite reduzir a carga computacional envolvida. As características devem conter toda a informação relevante do sinal áudio, sem perda dos dados relevantes. Foram implementadas duas abordagens: selecção através *método de histograma* e o *método SFFS* ("*Sequencial Forward Floating Selection*").

2.4. Classificação

Esta secção é composta por duas fases distintas: o treino do reconhecedor e a classificação.

2.4.1. Treino do reconhecedor (aprendizagem)

Para se realizar o treino é necessário, à priori, reunir um conjunto de pistas áudio que represente de forma o mais generalista possível, cada um dos instrumentos musicais.

Cada classe (instrumento) do conjunto de treino é constituída por um único ficheiro áudio, com duração máxima de 3 minutos (obtido através da concatenação de segmentos de 10s, no máximo, provenientes de

várias fontes). Após a extracção de características, o classificador tem por tarefa encontrar fronteiras de decisão (criação do modelo) que definam as regiões a que cada classe pertence; neste caso o classificador baseou-se num SVM binário ("*Support Vector Machine*").

Devido ao facto de este ser binário, o modelo obtido apenas caracteriza duas classes. Por conseguinte, para expandir a uma situação de multiclasse, foi utilizado o método de treino "*one versus one*" (determina-se todas as combinações binárias possíveis, gerando um modelo para cada par).

2.4.2. Classificação (reconhecimento)

Após a geração dos modelos, é possível agora submeter um ficheiro áudio genérico para determinar quais os instrumentos aí presentes. Para o problema de multiclasses, foi necessário implementar um algoritmo que não fosse muito dispendioso em termos de processamento (por causa do reconhecimento em tempo real), e sem grandes perdas em termos de robustez. A opção passou por se usar uma estrutura em forma de árvore em que cada um dos nós é um classificador binário; o *DDAG-SVM* [5]. Neste algoritmo, a decisão é "irreversível" (os nós que contenham a classe que foi preterida não entram mais no processo de classificação); por isso o reconhecimento é obtido após N-1 tomadas de decisão (para um total de N classes).

4. Apresentação de Resultados / Conclusão

		Oboé	***Trompete***	***Clarinete***	***Trompete***	***Baixo***	***Bateria***	***Sax***	***Piano***
ENTRADA	***Oboé***	98,7616	0	0,154799	0,309598	0,309598	0	0,154799	0,309598
	Trompete	0	98,7156	0	0,183486	0,183486	0,183486	0,366972	0,366972
	Clarinete	0	0,356	84,5005	5,2734	0,4839	3,7836	4,5667	1,0268
	Trompete	0,28329	0	0	64,0227	0,566572	1,55807	9,77337	23,796
	Baixo	0	0,959693	0	0,383877	89,4434	3,07102	0,575816	5,56622
	Bateria	0	3,1068	0	1,5534	25,6311	56,8932	6,99029	5,82524
	Sax	0	1,74603	0	24,7619	7,93651	2,69841	46,5079	16,3492
	Piano	0	0	0	9,10853	0	7,94574	18,9922	63,9535

Tabela1:Matriz de confusão para apenas 8 classes (em valores percentuais)

Da tabela 1, conclui-se que o reconhecedor tem um comportamento muito aceitável. No entanto convém salientar que a menor capacidade em discriminar o saxofone, e a bateria, é em parte devido à complexidade do conjunto de teste (demasiadamente "mascarados" com outros instrumentos), juntamente com uma menor generalização do conjunto de teste, uma vez que usando o método de validação cruzada, obteve-se uma precisão à volta dos 80%. O caminho a traçar para melhorar a performance, passa pelo aumento da generalização do conjunto de treino, bem como o uso de características que descrevam melhor o timbre.

5. Referências

[1]Moore, B. (1995). Hearing - Handbook of perception and cognition. Academic Press, Inc., London,
[2]Julius O. Smith, Jonathan S. Abel (1999) - Bark and ERB Bilinear Transforms. IEEE
[3] Klapuri, A., Eronen, A., and Astola, J. (2005). Analysis of the meter of acoustic musical signals. IEEE Trans. Speech and Audio Processing, in press.
[4] Juan Pablo Bello, Laurent Daudet. A Tutorial on Onset Detection in Music Signals.
[5] J. Platt, N. Cristianini. Large margin DAGs for multiclass classification. MIT Press, (2000).