Descrevendo o Som

Tiago Fernandes Tavares
UNICAMP

17/mar/2015

Desde eras muito anteriores aos computadores digitais, seres humanos já classificavam os objetos ao seu redor. Neste texto, argumento que métodos de classificação automática são apenas implementações da versão artesanal dessa arte. A computação digital permite ampliar o alcance da arte da classificação, mas o fator humano nunca deverá ser desconsiderado.

Inicialmente, discutirei na Seção 1 o que significa classificar um objeto. Nesta discussão, apresentarei os conceitos de descritores de alto nível – que têm significado ligado ao seu contexto – e de baixo nível – que podem ser medidos independente de seu contexto. Como será abordado, há uma diferença inevitável entre a informação que pode ser obtida de cada um desses tipos de descritores.

Na Seção 2, abordarei mais profundamente como descritores de áudio funcionam. De forma geral, descritores de baixo nível são inspirados em descritores de alto nível. Estes, por sua vez, vêm de estudos da musicologia e da psicoacústica.

Por fim, na Seção 3 discutirei implicações profundas dos problemas expostos por todo o texto. Abordarei brevemente evidências, tanto experimentais quanto dialéticas, de que o problema de descrição de áudio não pode ser resolvido de forma geral e ampla. Dessa forma, concluo que o problema de classificação estará sempre ligado ao contexto em que se insere, e, portanto, é um problema humano.

Esta discussão não pretende, de forma alguma, tornar-se um *landmark* científico. Porém, tenho a intenção de apresentar uma visão da questão da classificação de áudio que seja mais ligada à interdisciplinaridade com a qual trabalho. Trata-se, assim, de uma iniciativa no sentido de buscar uma linguagem que seja palatável para todos os interessados na área e, como efeito colateral, propõe uma bibliografia essencial para o estudo deste assunto.

1 Descritores

O problema de classificação consiste em atribuir rótulos a objetos [1]. Um objeto é uma coisa qualquer do mundo e um rótulo é uma forma simples de descrevê-lo. Essa descrição, através do rótulo, diferencia o objeto dos demais objetos existentes no universo em questão.

Tomemos como exemplo o caso dos dinossauros [2]. Dentre todo o universo de dinossauros, arbitrei que quatro espécies me interessam mais: tiranossauros, velocirraptors, triceratops e brontossauros. Um problema de classificação que pode usar essas espécies é o de, dado um dinossauro, atribuir-lhe o rótulo "espécie".

Para leitores aficcionados por dinossauros, essa definição é trivial, pois a aparência de cada um deles é bastante singular. No problema de classificação, porém, precisamos determinar motivos claros que guiem essa decisão, de forma que uma pessoa que nunca viu um dinossauro possa definir sua espécie. Uma possível solução para isso é construir uma tabela relacionando cada espécie de dinossauro a características que podem ser usadas como base para realizar a classificação, como mostrado na Tabela 1.

À partir dessa relação entre descritores e rótulos, podemos encontrar a espécie de nosso dinossauro. Um fóssil de um grande carnívoro bípede, por exemplo, deve ser identificado como um tiranossauro. Portanto, essa relação entre rótulos e descritores parece ser eficaz.

Porém, o problema de classificação de dinossauros

Espécie	Tamanho	Alimentação	Locomoção
Tirano.	Grande	Carnívoro	Bípede
Veloci.	Pequeno	Carnívoro	Bípede
Bronto.	Grande	Herbívoro	Quadrúpede
Tricer.	Pequeno	Herbívoro	Quadrúpede

Tabela 1: *Descritores de alto nível relacionados a dinossauros.*

que encontramos envolve medidas contínuas de fósseis. Não podemos nos esquecer que ser "grande" ou "pequeno" não é uma característica instrínseca do fóssil, mas o resultado de uma comparação com demais fósseis. Além disso, as características de alimentação e locomoção de fósseis não vêm da observação direta, mas sim de deduções feitas sobre as características diretamente observáveis de fósseis.

Num problema real, um biólogo irá encontrar dados mensuráveis quanto a um certo objeto, como mostrado na Tabela 2. Nesse caso, temos acesso ao comprimento de dois ossos de um certo fóssil, o fêmur e o antebraço. À partir dessas medidas, podemos inferir a espécie de dinossauro à qual esse osso pertence.

Osso	Comprimento
Fêmur	30 cm
Antebraço	5 cm

Tabela 2: *Descritores de baixo nível relacionados a um fóssil hipotético.*

Uma vez que o fêmur do fóssil tem 30 cm de comprimento, podemos inferir que se trata de um dinossauro pequeno. Além disso, o antebraço do fóssil é muito pequeno em relação ao fêmur, indicando que ele não é usado para locomoção. Assim, podemos inferir que se trata de um fóssil de um animal pequeno e bípede, e, de acordo com a Tabela 1, podemos atribuir-lhe o rótulo velocirraptor.

Neste processo de raciocínio, partimos de descritores de baixo nível, ou seja, diretamente mensuráveis, e construímos sua relação com descritores de alto nível, ligados a significados contextuais. Após essa construção, foi possível chegar ao rótulo que precisávamos. Esse passo, porém, pode conter erros de diversas naturezas.

Em nossa inferência, por exemplo, assumimos diretamente que o fóssil estudado pertencia a um adulto – um infante tiranossauro, por exemplo, poderia apresentar essas mesmas medidas. É muito comum que, em problemas de classificação, exista uma divergência sensível entre os descritores de baixo nível e os de alto nível. Essa divergência é chamada de *gap* semântico, e superá-la é um desafio significativo na área de áudio, como veremos a seguir.

2 Descritores de Áudio

A classificação de sinais de áudio tem diversas aplicações, incluindo a musicologia, a etnomusicologia e a didática musical. Seria difícil se referir a movimentos musicais sem utilizar rótulos como "andamento", "dinâmica" ou "grave" [3]. Esses rótulos emergem naturalmente em comunidades ligadas a determinadas culturas ou estilos musicais, e é comum que grupos diferentes utilizem rótulos diferentes para se referir ao mesmo fenômeno, ou mesmo que um único rótulo tenha signifcados diferentes em grupos diferentes [4].

Um interessante problema de classificação envolve diferenciar, através de descritores que podem ser entendidos por qualquer pessoa, a obra artística de dois compositores diferentes. Se tomamos dois compositores muito diferentes, como Beethoven e o contemporâneo Mr. Catra, esse problema terá uma certa dificuldade. Diferenciar a obra de Beethoven e de Mozart, porém, é um problema que exige um estudo mais cuidadoso, e podem surgir exemplos de compositores tão próximos quanto quisermos.

Para realizar essas tarefas, podemos aplicar descritores de alto nível, tais como "andamento" ou "instrumentação", e, em determinados contextos culturais (em especial, aqueles historicamente influenciados pela Europa Ocidental), pode ser interessante também estudar as relações entre os sons que formam a peça em questão [3]. Para tal, é possível utilizar notas musicais, exemplificados na Figura 1, que representam instruções para a execucão de ações de um repertório – frequentemente implicando na geração de sons específicos – em instantes de tempo pré-determinados. O significado exato dessas ações, porém, depende do contexto no qual a peça musical está inserida, ou seja, uma nota só se realiza ao ser executada por um intérprete e, portanto, notas musicais são descritores de alto nível.

Notas musicais permitem descrever peças de forma eficiente em determinados estilos, mas não são capazes de contemplar características como modificações de timbre, que podem ser importantes em vários contextos [5]. Por esse motivo, pode ser interessante utilizar descritores de baixo nível para auxiliar no processo de análise de uma peça. Para isso, verificaremos o que pode ser calculado sistematicamente, tomando por base gravações digitais, e que ainda traga informações úteis ao nosso processo de classificação.

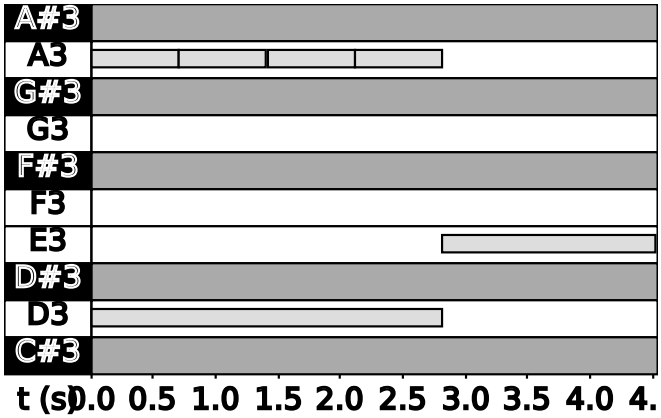

Figura 1: *Descrição em piano-roll de uma curta sequência de notas musicais. Nessa notação, o eixo horizontal representa o tempo, o eixo vertical representa a altura e as notas são marcadas com caixas em cada um dos corredores correspondentes.*

É comum que um sinal estacionário seja modelado como uma soma de sinais senoidais, cada um com sua amplitude, frequência e fase próprias[1]. Esse modelo é amplamente utilizado porque permite dividir o sistema estudado em partes mais fáceis de interpretar, sem que para isso qualquer informação sobre ele tenha sido destruída. Ainda, a Transformada Discreta de Fourier pode ser aplicada para calcular essa representação à partir de um sinal amostrado, tal qual um arquivo de áudio digital [6].

À partir de estudos psicoacústicos, sabemos que o ouvido humano é relativamente insensível à fase de sinais senoidais [7]. Devido a essa insensibilidade, dois assovios de mesma frequência e amplitude soam exatamente iguais. Isso indica que toda a informação auditiva em um trecho de áudio estacionário estará contida nas frequências e amplitudes que formam esse trecho.

Além disso, sabemos que as características perceptuais de sinais de áudio tendem a variar numa escala de tempo relativamente lenta, da ordem de décimos de segundo. Isso ocorre porque seres humanos são capazes de executar ações emissoras de som e perceber alterações em manifestações sonoras nessa escala de tempo, limitada pelas capacidades físicas do corpo [8]. Essa limitação se torna ainda mais evidente quando se considera a prática cultural de executar notas musicais, cuja duração é tipicamente superior a décimos de segundo.

Esse cenário inspira a estimação de espectrogramas, que são construídos dividindo um sinal de áudio em segmentos subsequentes de duração conhecida, chamados quadros. Calculando a Transformada Discreta de Fourier de cada um desses quadros, e descartando a informação de fase, temos uma série de elementos que relacionam tempo, frequência e amplitude. Esses elementos contém toda a informação que podemos receber auditivamente, e, portanto, podem ser usados para análise de áudio.

A Figura 2 mostra a representação de um espectrograma. Para cada ponto, temos três características: sua posição no eixo horizontal (tempo), no eixo vertical (frequência) e cor (cores mais escuras representam amplitudes maiores). Podemos visualizar que, em cada instante de tempo, há um conjunto de séries harmônicas presentes, e que esse conjunto varia com o tempo, criando assim uma progressão de notas.

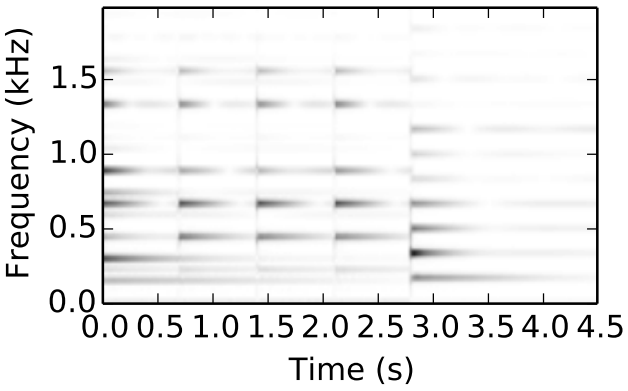

Figura 2: *Espectrograma extraído de uma curta sequência de notas de piano.*

À partir de todas essas discussões, é intuitivo pensar que sons parecidos terão espectrogramas parecidos. Porém, quantificar essa semelhança não é trivial. Para tal, é preciso abordar brevemente o campo da psicoacústica.

É sabido que determinadas características perceptuais do som são correlacionadas a características objetivas do sinal de áudio correspondente. Sabemos, por exemplo, que a sensação de intensidade sonora se correlaciona à potência instantânea de um sinal [9]. Assim, podemos calcular a energia desse mesmo sinal em cada quadro (e, portanto, sua potência naquele intervalo de tempo) e então visualizar uma curva que mostra a evolução da intensidade sonora ao longo de um trecho de áudio.

Nesse caso, temos um problema de escala, pois o significado de um valor qualquer não é necessariamente representativo ou mesmo fácil de se interpretar. A potência de um sinal sonoro pode ser descrita em decibéis (dB), o que parece corresponder mais linearmente à percepção auditiva, e é possível também encontrar faixas de frequência na qual a audição hu-

[1] Um sinal senoidal soa como um Teremin ou um assobio.

mana é, em média, mais sensível. Um refinamento ainda mais difícil diz respeito ao foco de atenção, ou seja, à capacidade humana de ouvir em mais detalhes uma determinada fonte sonora imersa numa mistura, em detrimento das demais fontes.

Mesmo para uma característica tão simples quanto a intensidade sonora, é difícil deduzir um modelo matemático que forneça uma representação precisa da percepção humana. Porém, isso não impede que o modelo forneça informações importantes para a análise de sinais sonoros. A Figura 3 mostra a energia de cada quadro da mesma gravação com a qual foi gerada a Figura 2, e nela pode-se verificar como picos de energia funcionam como indicadores do início de notas musicais

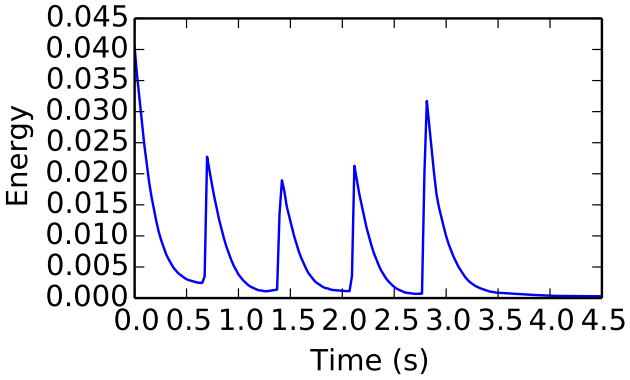

Figura 3: *Energia em cada quadro de uma curta sequência de notas de piano.*

Porém, a escala do eixo vertical dessa mesma figura evidencia que o significado numérico da energia não corresponde necessariamente à percepção de intensidade. Afinal, o que seria uma intensidade 0.2, ou 0.36? Nesse aspecto, é importante lembrar que sinais digitalizados de áudio já foram submetidos a uma série de equipamentos que podem ou não ter amplificado ou atenuado o sinal, de forma que um valor absoluto de intensidade no sinal gravado pode não corresponder ao mesmo valor no sinal acústico original.

Se nos afastarmos ainda mais da correspondência direta entre atributos perceptuais e descritores, podemos nos permitir a utilizar os coeficientes melcepstrais (MFCCs). Esses coeficientes são calculados aplicando uma função inspirada nas características de recepção da cóclea (as operações para isso serão omitidas). Como resultado, atribuímos um conjunto de coeficientes descritivos a cada quadro do espectrograma, e esses coeficientes fornecem uma descrição multidimensional do timbre ali contido [9].

MFCCs foram desenvolvidos com inspiração num modelo de emissão de voz no qual uma vibração, gerada pelas cordas vocais, atravessa um filtro modulador, correspondente ao restante do aparelho fonador. Realizando uma operação que aproxima o inverso de uma convolução, é possível recuperar as características do filtro modulador, e, portanto, do formato do aparelho fonador naquele momento. Essa informaçào permite estimar qual fonema está sendo emitido por uma pessoa, e, baseando-se nos bons resultados em reconhecimento de fala, MFCCs foram deliberadamente aplicados para o reconhecimento de outros tipos de sinais de áudio.

Embora MFCCs tenham sido utilizados em diversas aplicações úteis, eles não possuem correspondência psicoacústica. Isso implica que, para utilizá-los, é preciso utilizar métodos que descobrem automaticamente o mapeamento entre descritores e rótulos. Dentre esses métodos, encontramos os algoritmos de aprendizado de máquina, que estimam um possível mapeamento à partir de dados fornecidos como treinamento [10].

3 Discussões e conclusão

Neste texto, apresentei motivações para o uso de descritores de áudio para fins de análise. A discussão se iniciou mostrando que o problema de atribuir rótulos a áudio é semelhante ao de atribuir rótulos a qualquer objeto. Para o caso específico de sinais de áudio, porém, podemos aplicar conhecimentos oriundos da musicologia, psicoacústica e da matemática, de forma a construir uma base interdisciplinar de conhecimentos atuando juntos em prol da solução de um problema.

O *gap* semântico se mostra um problema importante quando utilizamos descritores. Isso fica claro quando utilizamos descritores de baixo nível, que funcionam para diversas aplicações, mas cujo ligação com características perceptuais pode ser apenas marginal ou mesmo inexistente. Lembro, porém, que há um *gap* semântico mesmo em descritores de alto nível.

Numa oficina realizada em nosso laboratório, formei um grupo com mais dois colegas e nos propusemos a cada um, individualmente, classificar trechos pré-definidos de uma peça musical. A peça em questão era a Nomos Alpha, de Xenakys, e cada trecho deveria ser classificado como “pontos” ou “linhas”. Verificamos que nós, mesmo tendo uma formação parecida, concordávamos com a classificação apenas em 60% dos trechos escolhidos.

Um dos motivos para isso foi a inconsistência da

definição de nossos rótulos. Porém, é importante salientar que isso aconteceu mesmo tendo o grupo definido os rótulos em conjunto. Um dos colegas chegou a ressaltar que alguns trechos seriam, de certa forma, pontos, mas, por um outro ângulo, seriam linhas, o que constitui uma ambiguidade perceptual que é muito comum e valiosa na música.

Num exemplo simples, tomemos uma nota dó tocada por um saxofone. Ela será mais parecida, perceptualmente, com a mesma nota, tocada por um piano, ou com uma nota ré, tocada pelo mesmo saxofone? Podemos argumentar em favor de ambas as hipóteses, e um bom arranjador saberá explorar esse tipo de ambiguidade criando novos elementos para a composição de uma peça.

Dois paradigmas são bastante comuns em análise musical (e de áudio, em geral). O primeiro deles assume que a música traz, em si, relações ontológicas, de forma que cabe ao ser humano descobri-las. O segundo paradigma pressupõe que a música apenas desperta, no ser humano, uma série de sensações que são percebidas como vindas da música [11].

De uma forma ou de outra, em algum momento será necessário definir o que queremos conseguir com uma certa análise. À partir disso, é possível definir um plano de execução e balisar a análise de resultados. Nesse processo, será sempre necessário considerar que há uma diferença entre a característica perceptual que gostaríamos de modelar e o descritor de baixo nível que somos capazes de calcular.

Neste texto, também discuti como essa diferença entre o que observamos e o que rotulamos é inevitável, ao menos nos casos de dinossauros e música. O mesmo pode valer para diversos outros casos de classificação. Portanto, sistemas que classificam objetos do mundo real inevitavelmente terão que lidar com falsos positivos e falsos negativos, que vêm de falhas existentes tanto em alto quanto em baixo nível.

Ainda, é possível que os próprios rótulos que escolhemos não tenham sentido para um grupo muito grande de pessoas. Gêneros musicais, por exemplo, são rótulos que podem servir para organizar coleções de gravações, mas, ao mesmo tempo, são bastante subjetivos. Isso é um reflexo do fato de que, embora certas manifestações musicais sejam claramente diferentes, a fronteira perceptual entre duas categorias é arbitrária e deriva de motivos extrínsecos à música.

Na verdade, podemos extrapolar esse mesmo problema para outros campos ligados à compreensão de fenômenos perceptuais. Um rótulo só tem significado se estiver ligado a um contexto, e esse contexto pode ser tão particular quanto se queira. Assim, não há uma maneira inerentemente mais adequada de se classificar objetos, já que a resposta mais adequada depende da aplicação.

Assim, faz-se valer a antiga frase: "todos os modelos estão errados, mas alguns são úteis" [12]. Em campos delicados, como a atribuição de rótulos a objetos, sempre estaremos sujeitos a erros, e isso deve ser levado em consideração. Por esse motivo, encerro este texto acreditando que, muitas vezes, o resultado de uma classificação será menos importante que o processo de análise e criação na qual ela foi inspirada.

Referências

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2 ed., October 2000.

[2] W. D. Mathew, *Dinossaurs, with special reference to the American Museum collections*. Amerian Museum of Natural History, 1915.

[3] A. Schoenberg, *Fundamentals of Musical Composition*. Faber and Faber Limited, 1 ed., 1970.

[4] C. Upton, B. Eaglestone, and N. Ford, "The compositional processes of electroacoustic composers: Contrasting perspectives," in *Proceedings of the ICMC*, 2005.

[5] D. Cope, *Techniques of the Contemporary Composer*. Schirmer Thomson Learning, 1997.

[6] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*. Prentice Hall Inc., 2 ed., 1999.

[7] H. Helmholtz, *On the Sensation of Tone*. Dover Publications Inc., 4 ed., 1885.

[8] H. F. Olson, *Music, Physics and Engineering*. Dover Publications Inc., 2 ed., 1967.

[9] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. Springer Science Business Media LLC, 2006.

[10] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2 ed., 2000.

[11] L. B. Meyer, *Emotion and Meaning in Music*. Chicago Press, 1956.

[12] G. E. P. Box and N. R. Draper, *Empirical Model Building and Response Surfaces*. John Wiley & Sons, 1987.