PERCEPÇÃO E REPRESENTAÇÃO VETORIAL DE TIMBRE

Dami Duarte

Departamento de Engenharia Elétrica, Universidade Federal de Sergipe - UFS
Campus de São Cristóvão - Sergipe, CEP. 49100-000 , Brasil

JUGURTA MONTALVÃO

Departamento de Engenharia Elétrica, Universidade Federal de Sergipe - UFS
Campus de São Cristóvão - Sergipe, CEP. 49100-000 , Brasil

E-mails: dami_doria@hotmail.com, jmontalvao@ufs.br

Abstract—In this paper, the minimum dimension required for the vector representation of timbres is analysed based on perception and subjective comparison of paired sounds, associated to multidimensional scaling of scores from opinions gathered through interviews. The results thus inferred (through multidimensional scaling) corroborate the dimensions commonly used in mel-cepstral coefficient representation of sounds, but both the analysis and collected subjective measure further provide new basis for advances in researches on new representation spaces.

Keywords—Timbre, Multidimensional Scaling, *Mean Opinion Score*.

Resumo—Neste artigo, a dimensão vetorial mínima necessária à representação de timbres é analisada com base na percepção e comparação subjetiva de sons pareados, aliadas ao uso do escalonamento multidimensional e entrevistas. A dimensão de representação é então inferida pela aplicação do escalonamento multidimensional. Os resultados obtidos corroboram as dimensões correntemente usadas em representações sonoras baseadas em coeficientes mel-cepstrais, mas tanto as análises quanto as medidas subjetivas coletadas também servem de base à pesquisa de novos espaços de representações.

Palavras-chave— Timbre, Escalonamento Multidimensional, *Mean Opinion Score*.

1. Introdução

O timbre acústico é um atributo perceptual do som, cuja representação formal não é evidente, o que dá margem ao surgimento de uma miríade de representações concorrentes, sendo que nenhuma delas pode ser considerada um consenso.

A maioria das definições de timbre o classificam pelo que ele não é, ou seja, ao analisar as características do som, certos aspectos são notados e descritos diretamente, como é o caso da duração e do volume, já a altura do som (chamada também de *pitch*) é um aspecto psicoacústico de difícil modelagem pois, embora esteja relacionado à percepção da frequência fundamental (F0) de um som periódico, o *pitch* não é um atributo físico do som, o que representa um problema de representação à parte. No entanto, mesmo com uma representação adequada da altura, da intensidade e da duração, ainda assim um som não fica completamente representado, e essa lacuna de representação induz à concepção do atributo complementar conhecido como timbre sonoro (Shiraishi, 2006; Houtsma, 1997; Handel, 1995; De Poli, 2003). Assim como o *pitch*, o timbre também é um atributo psicoacústico, isto é, dependente da percepção auditiva, mas sabe-se que diversos aspectos físicos influenciam sua caracterização, o que pode levar a uma representação multidimensional adequada (Houtsma, 1997; Grey, 1975). Infelizmente, não há consenso sobre quais variáveis físicas o influenciam (Hajda, 2007; Handel, 1995; Donnadieu, 2007).

Em geral, na a análise de timbres, os aspectos físicos mais usados são (Hajda, 2007; Grey, 1975; Donnadieu, 2007; Handel, 1995):

- Fluxo espectral;
- Centróide espectral;
- Irregularidade espectral;
- Distribuição da energia espectral;
- Variações temporais da onda – quantidade de sons harmônicos ou inarmônicos no ataque, velocidade de decaimento, evolução temporal de componentes espectrais;

A maioria dos estudos de timbre usa a abordagem perceptual em primeira estância. Assim, numa abordagem experimental, é necessária a preparação de um conjunto pré-definido de

amostras de sons sintéticos ou naturais, com timbres supostamente diversos. A partir disso, são coletados dados perceptuais, baseados em entrevistas, correspondentes à percepção de "distância" entre pares de sons. Vale notar que há estudos nos quais os entrevistados são profissionais na área da música, como é o caso do experimento de Grey, que resultou num espaço tridimensional para um conjunto de 16 timbres (Grey, 1975), e o de Charbonneau em 1981 (Donnadieu, 2007), como também há estudos nos quais os entrevistados não são especialistas, como é o caso de Wessel e Krumhansl, que chegaram a um espaço tridimensional para uma base de 21 timbres sintéticos (Handel, 1995). Em todo caso, o conjunto das distâncias obtidas a partir das opiniões de voluntários forma uma matriz individual de dissimilaridades, e a consolidação de todas as matrizes individuais provê, finalmente, uma única matriz média de dissimilaridades, que consolida uma *Mean Opinion Score* (MOS) para um conjunto de observadores. Quanto maior e mais diverso o conjunto de observadores, mais representativo o MOS tende a ser. Com base em um dado MOS (ou na matriz média de dissimilaridades que a representa), e usando as técnicas de escalonamento multidimensional (MDS) é possível se distribuir os timbres em um espaço métrico hipotético de representação. As dimensões dos espaços que melhor acomodam a MOS fornecem indicações da dimensão mínima necessária à representação dos timbres, embora a representação propriamente dita não seja revelada.

Pesquisas acerca do timbre tipicamente são feitas ou pelo viés da fala ou pelo viés da música, o que tem levado a conclusões bastante ricas (Allen, 1994). Porém como o timbre transcende os sinais de fala ou música, optamos, neste artigo, por uma análise de timbre de um ponto de vista mais geral.

Na seção 2, são descritos os sinais acústicos usados na montagem do MOS, bem como o protocolo de coleta das medidas individuais de dissimilaridade entre pares de amostras. Na seção 3, é discutido o método do escalonamento multidimensional e a abordagem do mesmo na detecção das dimensões para o espaço de timbre. Na seção 4, são expostos os resultados obtidos nessa pesquisa. Na seção 5 são discutidos os resultados.

2. Base de Sinais

Neste trabalho, optamos por usar um conjunto de sons sintetizados a partir da combinação linear de sinais harmônicos com envelopes, amplitudes, fases e frequências aleatórias. Como as frequências se distribuem num intervalo limitado porém contínuo, a probabilidade desses sinais possuírem divisores comuns é nula, evitando assim a síntese de sinais com *pitch*, enfatizando a característica do timbre.

A intenção dessa combinação linear é criar respostas ao impulso sintéticas que englobem, aleatoriamente, características modais diversas. Isto é, para sistemas que modelam estruturas vibrantes simples (com apenas um modo de reverberação), como os ressonadores de Helmholtz, tais respostas têm a forma de uma senóide com decaimento exponencial. Para que o som adquira características quiméricas, de estruturas assimétricas com vários modos de reverberação, podemos combinar vários modos num mesmo sinal $s(n)$, de acordo com

$$s(n) = \sum_{i=1}^{10} A(i)\cos\left(2\pi \frac{f_0(i)}{F_a} n + \theta(i)\right) e^{-\alpha n} \qquad (1)$$

em que n representa um contador de amostras (tempo discreto) $n = 0, 1, 2, 3, \ldots, 5000$, numa taxa de $F_a = 8000$ amostras por segundo. Os parâmetros aleatórios de cada um dos10 modos componentes de cada timbre são a amplitude, A, escolhida aleatoriamente e uniformemente ente -0,5 e 0,5, a frequência central do modo, f_0, escolhida aleatoriamente e uniformemente entre 0 e 4000 Hz, a fase, θ, escolhida aleatoriamente e uniformemente entre 0 e 2π, e o decaimento, α, escolhida aleatoriamente e uniformemente entre 0,003 e 0,01.

Embora nossa escolha por sons sintéticos (em detrimento de sons naturais) nos permita criar um conjunto tão diverso quanto possível de sons, uma condição paradoxal se apresenta: como medir a diversidade dos timbres escolhidos para o experimento se o que buscamos são justamente as ferramentas apropriadas a essa medição? Como solução *ad hoc* ao impasse, fizemos uma primeira triagem subjetiva via análise perceptual de diversos sons produzidos dessa forma, com a ajuda do ouvido treinado de um músico, separando assim um subconjunto de 19 sons que acreditamos se espalharem satisfatoriamente por qualquer espaço razoável de representação de timbres. Além desses 19 sons, foram escolhidos mais 2 que, dada a forma

de síntese apresentada na fórmula (1), representam o que chamaremos de “sons opostos”, pois todos os parâmetros assumem valores extremos e opostos, respectivamente, para esse dois sons.

O protocolo de entrevistas foi baseado em um programa computacional com interface gráfica simplificada que escolhe, de forma aleatória, pares de sons e os apresenta ao observador. Além dos pares aleatórios de sons, cada observador escutou dois pares adicionais de sons: um par de sons iguais, e o par dos ditos “sons opostos”. Essa particularidade do protocolo de coletas de opiniões (*scores*) permitiu a análise preliminar da consistência das distâncias atribuídas por cada observador, provendo medidas de "fundo de escala". Cada entrevistado foi instruído a escutar quantas vezes quisesse cada som, pelo simples acionamento de botões, como ilustrado na figura 1. Após a escuta, o convidado foi instruído a ranquear a “diferença” entre os sons escutados numa escala com onze níveis de dissimilaridade, ou distâncias perceptuais. A figura 1 ilustra a interface gráfica do programa de entrevistas. Para facilitar a escolha das distâncias entre sons pelos entrevistados, foram atribuídas características perceptuais às opções, a opção zero, por exemplo, deveria ser atribuída a timbres considerados idênticos, enquanto a opção oito foi atribuída à timbres diferentes.

Figura 1: A interface do programa de entrevistas tem três botões, dois para tocar cada som da avaliação e um para gerar outro par de sons, e onze botões de opções no painel “Grau de Dissimilaridade”.

3. Análise de Escalonamento Multidimensional

Para se construir um espaço hipotético de timbres, utilizando o MDS, primeiro é necessário a construção de uma matriz que informa as distâncias entre pares de sons. Como usamos 21 (19+2) sons, essa matriz será quadrada, de ordem 21. A medida de distância de timbre é perceptual e é obtida por meio de entrevistas. Cada entrevistado é convidado a opinar sobre a distância relativa de dez pares de sons sorteados, numa gama de opções que vai de iguais (distância mínima, definida como zero) a opostos (distância máxima, definida como dez). Adicionalmente, além dos 10 pares sorteados, uma 11ª comparação é sistematicamente incluída (como estratégia metodológica para verificar o nível de atenção e percepção do entrevistado, bem como estabelecer uma referência relativa de "fundo de escala" por observador). Esse par adicional corresponde a sons com parâmetros extremos opostos. A figura 2 ilustra dois sinais da base de sons.

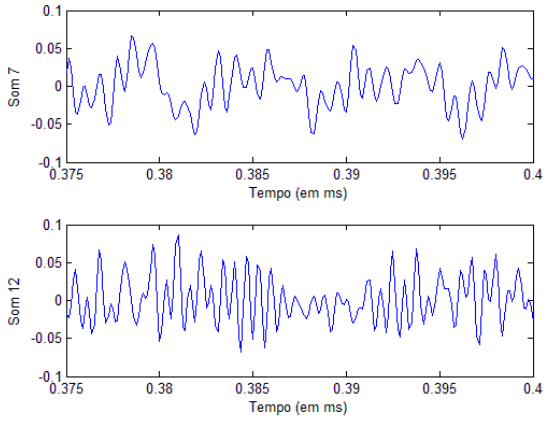

Figura 2: Ilustração dos sinais correspondentes ao “Som 7” e “Som 12”.

A matriz de dissimilaridades consolidada é gerada a partir de uma reformatação da matriz de dissimilaridades original devido a dois motivos principais: (1) para aplicar o escalonamento multidimensional à matriz de dissimilaridades é necessário que ela seja uma matriz simétrica com diagonal principal nula. A matriz de dissimilaridades original não é necessariamente simétrica, pois há uma flutuação de valores na comparação de pares iguais. Isto é, por exemplo, ao comparar o “som 7” ao “som 13” e, simetricamente, o “som 13” ao “som 7”, um dado entrevistado pode fornecer medidas de dissimilaridade diferentes. Desse modo, assumindo que essas flutuações são simétricas, as entradas da matriz são simetrizadas pela substituição de valores simétricos e discrepantes pela média ponderada desses valores. Adicionalmente, os valores da diagonal principal são forçados a zero. No entanto, esses valores são importantes para a análise da

consistência das avaliações, em que valores muito distantes de zero são considerados indícios de desatenção do entrevistado, e consequentemente todos os dados oferecidos pelo entrevistado serão descartados. A figura 3 representa, na forma de um histograma bi-dimensional, a matriz de dissimilaridades consolidada (mescla de todas as matrizes parciais), enquanto a figura 4 representa o os valores de média e desvio padrão para comparação de sons iguais.

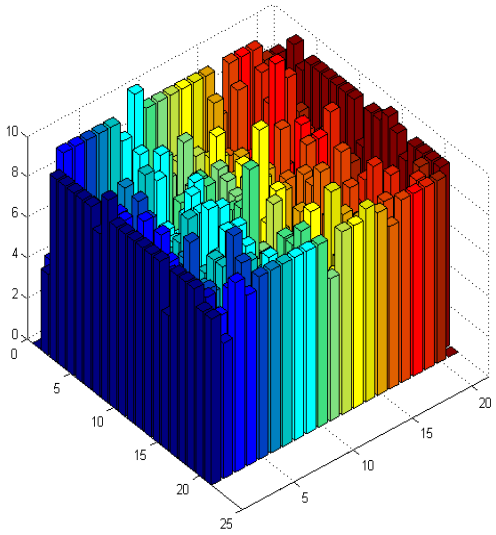

Figura 3: Matriz de dissimilaridades consolidada. É importante pontuar que a matriz de dissimilaridades é simétrica e com termos nulos na diagonal principal, condições básicas do MDS.

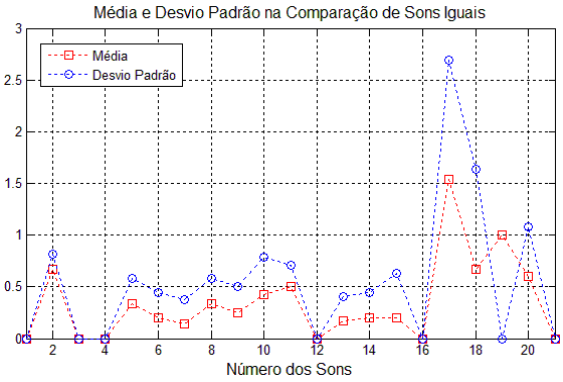

Figura 4: Valores de média e desvio padrão na comparação de sons iguais. De maneira geral, os valores se comportam bem, com possível exceção nos valores 17, 18 e 20, em que o desvio padrão sobe consideravelmente.

O MDS é um conjunto de técnicas de estatística para analisar dimensão de espaços de representação a partir de medidas de similaridades e dissimilaridades entre dados. Inicialmente, o MDS foi usado no campo da psicologia conhecido como psicometria (conjunto de técnicas utilizadas para mensurar, de forma adequada e comprovada experimentalmente, um conjunto ou uma gama de comportamentos que se deseja conhecer melhor) (Groenen, 2004; Rosman, 2008). Visto desse modo, a psicometria se encaixa perfeitamente nos objetivos desse estudo, sendo o MDS a ferramenta ideal para análise. Dentre os vários tipos de MDS, o mais interessante para ser utilizado nesta pesquisa é o chamado MDS métrico, que consegue distribuir uma série de pontos (timbres, no nosso caso) em um número de dimensões arbitrariamente escolhido, somente utilizando uma matriz de dissimilaridades. O MDS métrico se baseia em um algoritmo que garante o decaimento monótono da função de perda, ou *stress*, que é o quadrado do somatório das diferenças entre cada termo da matriz de dissimilaridades e cada termo da matriz de distância dos pontos distribuídos. A função de stress pode ser escrita como

$$s(\boldsymbol{X}) = \sum_{i<j} \omega_{ij}\left(d_{ij}(\boldsymbol{X}) - \delta_{ij}\right)^2 \qquad (2)$$

em que:

- **X** é a matriz *n* (número de pontos) por *p* (dimensões) de pontos distribuídos no espaço de *p* dimensões;
- ω é o peso relacionado;
- d_{ij}(X) é a distância entre os pontos da linha *i* e *j*;
- δ_{ij} é o termo da linha *i* coluna *j* da matriz de dissimilaridades Δ.

4. Resultados

Para a análise de escalonamento, a matriz de dissimilaridades consolida um total de 760 avaliações, numa média de 1,72 avaliação para cada par de sons, com um desvio padrão médio de avaliação de 0,8285. É importante mostrar que a diversidade de pessoas entrevistadas foi um aspecto desejável deste trabalho. Além disso, os entrevistados não conheciam o objetivo final da coleta de dados, isto é, não foram influenciados por esse objetivo, o que poderia levar a uma atenção mais focada em determinadas características do som. Ao contrário, buscou-se uma postura de escuta reação mais espontâneas possíveis.

Como esperado, o *stress*, na análise MDS, decai monotonicamente com o acréscimo de dimensões ao espaço de representação dos timbres. Além disso, é observado que, a partir de uma dada dimensão, essa medida de *stress* se "estabiliza". Isto

é, além dessa dimensão limítrofe, o MDS não mais consegue reorganizar os pontos numa geometria que diminua notavelmente o *stress*. Quando a estabilização do *stress* é percebida, pela análise dos gráficos de dimensão versus *stress*, podemos dizer que a dimensão mínima de representação do espaço vetorial de timbres - alvo da pesquisa - foi encontrada.

Variando o número de dimensões de 1 a 16, como apresentado na figura 4. Nota-se que, para o conjunto de timbres estudado, essa dimensão mínima é igual a 13, para a qual o *stress* estabiliza em um valor muito próximo de zero, indicando que o uso de dimensões adicionais seria redundante. Vale notar que, como a análise com MDS foi feita a partir de uma base de 21 sons, conseguimos analisar adequadamente o stress até a 20ª dimensão, e que seria necessário acrescentar novos timbres (e novos scores de entrevistas) à base caso a dimensão mínima não fosse encontrada nesses primeiros experimentos.

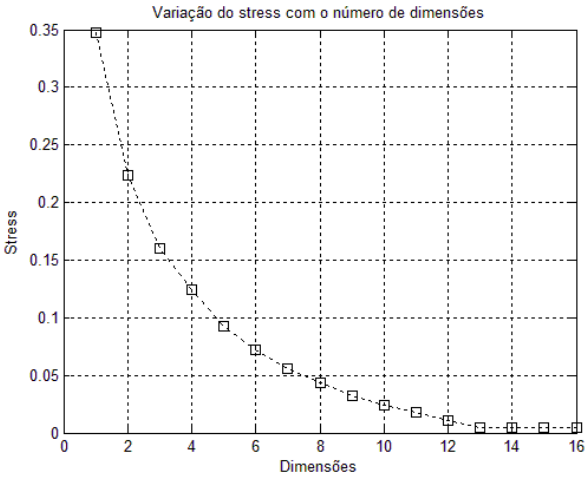

Figura 4: Gráfico *stress* versus dimensão. A partir da dimensão 13 o *stress* estabiliza.

4. Conclusões

O artigo apresenta os resultados de pesquisa da dimensão mínima do espaço de representação vetorial de timbres, em um procedimento experimental baseado na articulação de entrevistas (76 entrevistados), MOS (760 scores coletados) e análise de dimensões via MDS, resultando num espaço de 13 dimensões para a representação do timbre. É notável que trabalhos concorrentes sobre o mesmo tema ainda não tenham chegado a um consenso (Houtsma, 1997, Grey, 1975; Handel, 1995; Siedenburg, 2009). No entanto, existe um paralelo interessante com o valor aqui encontrado, que corrobora a validade da dimensão que estimamos como a mínima, a saber: Em pesquisas que discutem objetos métricos para a representação da fala (com seus atributos altamente timbrísticos) e do próprio timbre, ao utilizar os coeficientes cepstrais de frequência mel (MFCC), nota-se uma preferência pelo uso de apenas 13 coeficientes (Terasawa et al., 2005; Logan, 2000), isto é, pela representação dos timbres de voz em apenas 13 dimensões. Talvez seja apenas uma coincidência, mas nós acreditamos que seja mais que isso, pois os MFCC representam um modelo computacional da percepção psicoacústica dos sons, incorporando, por exemplo, os conceitos de banda crítica e escalas logarítmicas de percepção. Em todo caso, é digno de atenção o fato de que duas diferentes abordagens do timbre, uma via entrevistas perceptuais e MDS (a realizada aqui) e outra via resultante do ajuste, ao longo das décadas, das ferramentas de processamento de voz, chegam ao mesmo resultado em termos de dimensão preferencial para representação vetorial.

Agradecimentos

Este trabalho contou com o apoio financeiro do programa de inclusão em iniciação científica (PIIC) da Universidade Federal de Sergipe, e do CNPq. Agradecemos também aos voluntários que participaram das sessões de coleta de MOS.

BIBLIOGRAFIA

Referências

Allen, J. B. (1994). *How Do Humans Process and Recognize Speech, IEEE Transactions on Speech and Audio Processing, Vol. 2*.

De Poli, H. J. G. (2003). *"Perception Of Attributes In Real And Synthetic String Instrument Sounds", Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing.*

Donnadieu, S. (2007). *Mental Representation of the Timbre of Complex Sounds in the Analysis, Synthesis, and Perception of Musical Sounds. University of Illinois at Urbana.*

Grey, J. M. (1975). *Multidimensional Perceptual Scaling of Musical Timbres, Center for Computer Research in Music and Acoustics, Department of Music, Stanford, California.*

Groenen, P. J. F., Velden, M. (2004). *Multidimensional Scaling, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.*

Hajda, J. M. (2007). *The Effect of Dynamic Acoustical Features on Musical Timbre in the Analysis, Synthesis, and Perception of Musical Sound, University of Illinois at Urbana.*

Handel, S. (1995). *Timbre Perception and Auditory Object Identification in Hearing. Academic Press.*

Houtsma, A. J. M. (1997). *Pitch and Timbre: Definition, Meaning and Use, Journal of New Music Research*.

Logan, B. (2000). *Mel Frequency Cepstral Coefficients for Music Modeling, Cambridge Research Laboratory, Cambridge.*

Rosman, G. (2008). *Efficient Flattening in Manifold Learning and Image Processing, Senate of the Technion, Israel Institute of Technology, Haifa.*

Shiraishi, S. (2006). *A Real-Time Timbre Tracking Model Based on Similarity, The Hague, Royal Conservatory*.

Siedenburg, K. (2009). *An Exploration of Real-Time Visualisations os Musical Timbre, The Center for New Music and Audio Technologies University of California, Berkeley.*

Terasawa, H., Slaney, M., Berger, J. (2005). *A Timbre Space for Speech in Interspeech. Center for Computer Research in Music and Acoustics, Stanford University, Stanford, California*.

Terasawa, H., Slaney, M., Berger, J. (2005). *Percptual Distance in Timbre Space, Eleventh Meeting of the International Conference on Auditory Display, Limerick, Ireland*.