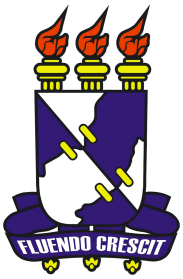
Dami Dória Narayana Duarte

Um estudo da relevância da dinâmica espectral na classificação de sons domésticos

São Cristóvão – SE, Brasil

Fevereiro de 2016



Um estudo da relevância da dinâmica espectral na classificação de sons domésticos

Dami Dória Narayana Duarte

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica – PROEE, da Universidade Federal de Sergipe, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Jugurta Rosa Montalvão Filho

São Cristóvão – SE, Brasil

Fevereiro de 2016

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Duarte, Dami Dória Narayana
D812e Um estudo da relevância da dinâmica espectral na classificação de sons domésticos / Dami Dória Narayana Duarte ; orientador Jugurta Rosa Montalvão Filho. – São Cristóvão, 2016.
67 f. : il.

Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Sergipe, 2016.

1. Som - Classificação. 2. Ruído - Medição. 3. Ondas sonoras. 4. Markov, espectros de. I. Montalvão Filho, Jugurta Rosa. II. Título.

CDU 534.61

Dami Dória Narayana Duarte

Um estudo da relevância da dinâmica espectral na classificação de sons domésticos

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica – PROEE, da Universidade Federal de Sergipe, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

São Cristóvão – SE, Brasil, 19 de fevereiro de 2016:

Prof. Dr. Tiago Fernandes Tavares
Convidado 1

Prof. Dr. Leonardo Nogueira Matos
Convidado 2

Prof. Dr. Jânio Coutinho Canuto
Convidado 3

Prof. Dr. Hendrik Teixeira Macedo
Convidado 4

São Cristóvão – SE, Brasil
Fevereiro de 2016

RESUMO

Este trabalho é um estudo da característica da dinâmica espectral em sinais sonoros, com vistas a encontrar as regularidades que podem ser modeladas em sons tipicamente domésticos, com o objetivo de classificá-los. O ponto de partida é o trabalho de Sehili et al. [1], no qual é proposto um sistema de classificação de sons domésticos baseado em GMM. O sistema de Sehili é reproduzido neste trabalho como marco zero na análise da dinâmica espectral, seguindo o mesmo roteiro dos experimentos. A partir daí, três conjuntos de experimentos são realizados, organizados de forma que, a cada novo experimento, uma técnica – que destaca um aspecto diferente da dinâmica espectral – seja incorporada. A primeira técnica analisada é a inserção da informação de gradiente discreto dos vetores de características, estratégia que representa uma análise de dinâmica espectral local e que resulta num aumento perceptível na taxa de classificação. O próximo experimento é realizado com um classificador baseado em HMM, no qual a informação de dinâmica espectral deve ser codificada na matriz de probabilidades de transição de estados do modelo. Os testes com o HMM não resultam em melhora na taxa de reconhecimento das classes de sons. O último experimento é baseado num extrator de características proposto pelo autor, chamado de Padrões de Envelopes de Energia por Banda (PEEB). O PEEB é um extrator que destaca os padrões de evolução espectro-temporais do sinais. Nos testes de reconhecimento de sons domésticos, o sistema de classificação baseado numa combinação das estratégias PEEB, MFCC e GMM resultam numa melhora significativa em relação a todos os outros sistemas testados. Conclui-se, com base nos resultados, que a dinâmica espectral dos sinais da base estudada é relevante à tarefa de classificação. No entanto, as maneiras de extração da informação de dinâmica espectral estudadas neste trabalho não são definitivas, pois ainda há muito espaço para desenvolvê-las. Por exemplo, no caso do PEEB, nota-se que a taxa de classificação fortemente é dependente da classe sonora, sugerindo formas mais elaboradas de fusão das características PEEB e MFCC para cada classe.

Palavras-chaves: sons domésticos, GMM, MFCC, delta MFCC, dinâmica espectral, monitoramento doméstico, HMM, simbolização, PEEB.

ABSTRACT

This work presents a study of the spectral dynamics characteristics of audio signals. More specifically, we aim at detecting regularities that can be modeled in typical domestic sounds, in order to classify them. Our starting point is the work of Sehili et al. [2], in which a household sounds classification system based on GMM is proposed. The Sehili system is reproduced in this work as a baseline system. Following the same protocol of experiments, a 73 % recognition rate is achieved. Afterwards, three sets of experiments are performed, arranged so that each new approach incorporates a new technique to highlight a different aspect of the spectral dynamics. The first technique is the insertion of the discrete gradient information of feature vectors, a strategy aimed at a local spectral dynamic analysis, and resultes in a perceptible increase in recognition rate. The next experiment is conducted with a HMM based classifier, in which the spectral dynamic should be encoded in state transition probability matrices. The tests with the HMM do not result in improved recognition rates. The last experiment is based on a features extraction method, proposed by the author, called Patterns of Energy Envelope per Band (PEEB). The PEEB is an extractor that highlight the signal spectral dynamics inside narrow bands. In domestic sounds recognition tests, the classification system based on a combination of PEEB, MFCC and GMM strategies resulted in a significant improvement over all other systems tested. We conclude, based on our results, that the spectral dynamics of the studied dataset plays an important role in the classification task. However, the approaches for spectral dynamic information extraction, studied in this work, are not definitive, for it is clear that they can be further developed. For example, in the case of PEEB, the recognition rate is strongly dependent on the sound class, suggesting more elaborate forms of fusion of PEEB and MFCC features for each class.

Key-words: domestic sounds, GMM, MFCC, delta MFCC, spectral dynamics, daily sound recognition, audio surveillance, HMM, PEEB.

LISTA DE ILUSTRAÇÕES

Figura 1 – Gráfico da população nas faixas de idade de 0-4 anos, 0-14 anos e 60 anos ou mais, no período de 1950 até 2010 (projeção até 2050). Figura retirada do relatório de envelhecimento do Fundo de Populações das Nações Unidas [2]. 12
Figura 2 – Espectrogramas de sinais corriqueiros em sistemas automáticos de monitoramento e televigilância de áudio. 14
Figura 3 – Um exemplo da quantidade de informação que áreas de mudanças abruptas do sinal carregam. Na imagem 3b é possível reconhecer o texto e o formato dos objetos apenas sinalizando pequenas áreas que representam bordas. 15
Figura 4 – A diferença da quantidade de informação visual que cada tipo de representação de um sinal sonoro carrega. Sabendo que esse sinal é uma elocução da palavra "aranha", na representação tradicional, só é possível localizar uma área de mudança rápida das características do sinal e podemos atribuir isso à elocução da consoante 'R'. Já na representação correspondente ao espectrograma, se percebem diversas nuances, como a presença de formantes e de suas modulações. 16
Figura 5 – Trecho do Soneto da Separação de Vinícius de Moraes e sua reprodução com supressão de consoantes e vogais (Figura 5b e 5c, respectivamente). O texto com supressão de vogais ainda conta com algumas vogais para não suprimir palavras de uma letra só. Um exemplo de que consoantes transmitem mais informações que vogais em texto escrito. 17
Figura 6 – Espectrograma da elocução da palavra 'aranha' à esquerda e, à direita, o espectrograma construído a partir dos padrões de modulação de Dudley (aqui foram usados 20 canais para uma melhor apresentação da figura, embora apenas 10 canais sejam suficientes, segundo Dudley). Ambas figuras representam frequências, na faixa de 0 a 4000 Hz, na direção vertical e o tempo no eixo horizontal. 19
Figura 7 – Um sinal sintetizado a partir de um seno a 250 Hz e seu primeiro harmônico. Após um intervalo de tempo, o harmônico muda em 180º sua fase, gerando uma região abrupta de transição (que pode ser notada como uma consoante). Escutando em um fone de ouvido, a maioria das pessoas percebe a mudança de fase. 21
Figura 8 – Rotina de extração de características baseada em suavização do espectro de um sinal de áudio. 25

Figura 9 – A etapa de classificação. 27
Figura 10 – Uma rede de Markov com três estados $\{S_1, S_2, S_3\}$ e nove probabilidades de transições entre estados $\{a_{11}, a_{12}, ..., a_{33}\}$. 30
Figura 11 – Erro de classificação em função do limiar de decisão no problema de discriminar Música e Fala, usando como característica apenas o desvio padrão da série temporal de um coeficiente MFCC. 37
Figura 12 – Primeira estratégia de classificação de Sehili et al. [1], baseada em GMM e MFCC. 38
Figura 13 – Taxas de reconhecimento na reprodução do trabalho de Sehili et al. [1], a partir de três tipos de inicializações diferentes do GMM. 40
Figura 14 – Ilustração do ganho de informação ao incorporar variação temporal aos vetores de características. 42
Figura 15 – Ilustração da diferença na estimativa do coeficiente de variação do vetor da posição t em relação às equações 3.1 e 3.3, para $k = 4$. 44
Figura 16 – Resultado dos testes com ΔMFCC a partir da eq. 3.1. 45
Figura 17 – Resultado dos testes com ΔMFCC a partir da eq. 3.3. 45
Figura 18 – Desempenho do sistema em função de k com coeficiente delta calculado a partir da eq. 3.1 (linha tracejada) e da eq. 3.3 (linha contínua). . . . 46
Figura 19 – Taxas de reconhecimento com classificador baseado em HMM sem treinamento em comparação com o classificador baseado em GMM. . . 50
Figura 20 – Variação da verossimilhança média dos modelos das 18 classes em função do número de iterações na etapa de treinamento. 51
Figura 21 – Algumas matrizes de probabilidades de transição A após o treinamento com o método de Baum-Welch. A característica comum é a concentração de probabilidade na diagonal principal. 51
Figura 22 – O traço da matriz A dividido por 30 em função das 18 classes de sons. Linha contínua para matrizes com treinamento tradicional e linha tracejada para matrizes com treinamento a partir de vetores de características subamostrados. 52
Figura 23 – Dez filtros passa-faixa com os centros espaçados igualmente em escala Mel. 55
Figura 24 – Esquema do extrator de características baseado em simbolização do perfil de energia por banda do sinal. 56
Figura 25 – Taxa de reconhecimento em função do coeficiente de ponderação ρ da combinação de *scores*. 58
Figura 26 – Comparativos dos sistemas de reconhecimento de sons domésticos testados ao longo deste capítulo. 59

LISTA DE TABELAS

Tabela 1 – Descrição das classes da base de dados. 34
Tabela 2 – Matriz de confusão com taxa de reconhecimento de 74,2%. As classes são rotuladas de acordo com a tabela 1. 41
Tabela 3 – Resultado dos testes com $\Delta\Delta$MFCC. 46
Tabela 4 – Matriz de confusão com taxa de reconhecimento de 77,1%. As classes são rotuladas de acordo com a tabela 1. 47
Tabela 5 – Taxas de reconhecimento com classificador baseado em HMM com treinamento da matriz A. 50
Tabela 6 – Matriz de confusão do esquema de classificação baseado em PEEB. Taxa de reconhecimento de 55,9%. 57
Tabela 7 – Taxas de reconhecimento com classificador baseado na combinação das técnicas do MFCC, GMM e PEEB, com $\rho = 0,75$. 58

LISTA DE ABREVIATURAS E SIGLAS

MFCC	*Mel-Frequency Cepstral Coefficients*
GMM	*Gaussian Mixture Model*
SVM	*Support Vector Machine*
RASTA	*RelAtive SpecTrAl*
DWTC	*Discrete Wavelet Transform Coefficient*
HMM	*Hidden Markov Model*
EM	*Expectation Maximization*
LPC	*Linear Predictive Coding*
PLP	*Perceptual Linear Prediction*
PDF	*Probability Density Function*
PMF	*Probability Mass Function*
FIR	*Finite Impulse Response*
PEEB	Padrões de Envelope de Energia por Banda

SUMÁRIO

1 Introdução 11
1.1 Monitoramento doméstico e televigilância sonora 11
1.2 A dinâmica espectral em sinais sonoros 15
1.3 Organização do trabalho 22

2 Aspectos gerais do reconhecimento automático de sons e o patamar inicial da pesquisa 23
2.1 Classificadores baseados em Mistura de Gaussianas e Modelo Oculto de Markov 26
2.2 Classificação de sons domésticos com GMM 33

3 Resultados Experimentais 36
3.1 Roteiro de pesquisa 36
3.2 Esquema de classificação básico, reproduzido de Sehili et al. [1] 38
3.3 Inserção da informação de variação do MFCC 42
3.3.1 Testes da influência do ΔMFCC na tarefa de classificação dos sons . . . 44
3.4 Sistema com classificador baseado em Modelo Oculto de Markov 47
3.5 Padrões de Envelope de Energia por Banda 53

4 Conclusões 60
4.1 Trabalhos futuros 62

Referências 63

1 INTRODUÇÃO

O estudo do reconhecimento automático de som esteve por muito tempo atrelado ao reconhecimento de fala/orador [1], afinal de contas, não são tarefas simples de serem executadas. Apesar de muitas técnicas importantes terem sido desenvolvidas ao longo de décadas de estudo, o reconhecimento de sons além da fala humana foi deixado de lado até pelo menos as últimas duas décadas [3], quando as modalidades de monitoramento automático de sons domésticos e vigilância baseada em áudio entraram em evidência.

Este trabalho tem como objetivo fazer um estudo das características dos sinais que são englobados nas modalidades de monitoramento doméstico e vigilância de áudio, sons notavelmente diversos, na mesma proporção da própria cognição auditiva humana. O aspecto mais importante neste estudo é a dinâmica espectral, característica que é reconhecidamente fundamental para a percepção e distinção de sons (desde pelo menos 1940 [4]). Apesar disso, a informação que a dinâmica espectral carrega não é facilmente capturada, principalmente quando se analisam os métodos de extração de características mais 'famosos' na literatura. Não obstante, a dinâmica espectral tem sido usada como argumento para a introdução de novos métodos no reconhecimento de som, como aqueles baseados em gradiente dos vetores de características (e.g. delta MFCC, delta delta MFCC), ou nos Modelos Ocultos de Markov (HMM). De fato, desde a década de 1990, tem havido um interesse crescente no estudo da dinâmica espectral como portadora da informação na fala [5].

Sem perder de vista o contexto mais geral do estudo da dinâmica espectral, este trabalho usa como patamar inicial o artigo "Daily Sound Recognition Using A Combination Of GMM And SVM For Home Automation" de Sehili et al. [1], ao qual acrescentamos novas análises (pelo viés da dinâmica espectral), mas mantendo a mesma base de sons domésticos. Para tanto, os métodos e experimentos sugeridos por Sehili et al. [1] foram reproduzidos em detalhes (ao ponto da recodificação em nova linguagem computacional) e, a partir daí, diversas técnicas foram testadas no intuito de incorporar a informação da dinâmica espectral no sistema de classificação, culminando no desenvolvimento de um extrator de características baseado nos conceitos de Dudley [4].

1.1 Monitoramento doméstico e televigilância sonora

A população idosa vem crescendo mais rapidamente que o total da população desde, pelo menos, a década de 1950 [2]. Nos EUA, por exemplo, a população com 75 anos ou

[1] Com algumas exceções, como é o caso de trabalhos voltados a sons musicais, por exemplo.

mais cresceu a uma taxa de 2,8%, enquanto a população total cresceu a uma taxa de 1,2% no período de 1950 a 2006 [2]. No Brasil, o número de idosos dobrou no período de 2001 a 2011, de acordo com o Relatório de Envelhecimento no Brasil da Secretaria de Direitos Humanos [6]. "Uma em cada 9 pessoas no mundo tem 60 anos ou mais, e estima-se um crescimento para 1 em cada 5 por volta de 2050", afirma o Fundo de População das Nações Unidas, "[...] Em 2050, pela primeira vez, haverá mais idosos que crianças menores de 15 anos" [7] (ver Figura 1). Esta realidade tem gerado uma onda de pesquisa científica e tecnológica para monitoramento domiciliar, a domótica, com ênfase na população idosa. Apesar dessa tendência não ser recente (o conceito *domotique* foi introduzido inicialmente na França, na década de 1980), só com os avanços tecnológicos dos últimos anos — como o barateamento do custo de microcontroladores, o desenvolvimento da internet — é que a domótica está mais presente no dia-a-dia [8].

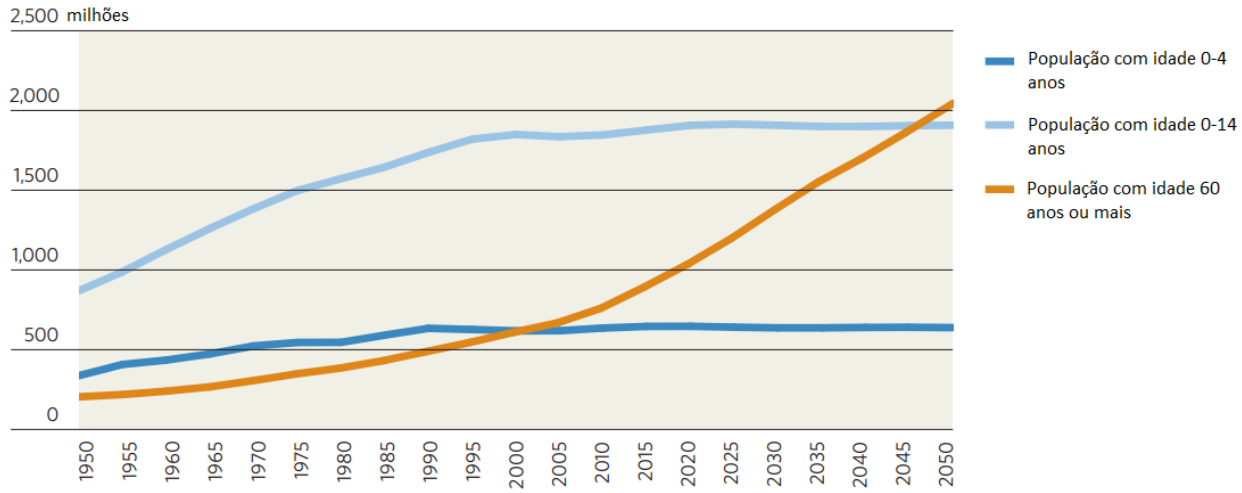


Figura 1 – Gráfico da população nas faixas de idade de 0-4 anos, 0-14 anos e 60 anos ou mais, no período de 1950 até 2010 (projeção até 2050). Figura retirada do relatório de envelhecimento do Fundo de Populações das Nações Unidas [2].

Dentre os sistemas de monitoramento doméstico desenvolvidos atualmente, se destacam os que utilizam sensores infravermelhos para mapear a movimentação de indivíduos [9], a utilização de telefone e internet para assistência remota [8, 10], câmeras para reconhecimento de imagem, microfones para reconhecimento de som e fala [1, 11–13] e ferramentas mais intrusivas como medidores de pressão arterial, glicemia etc [14].

O reconhecimento de som para monitoramento e televigilância é uma das modalidades mais promissoras na área de reconhecimento de som [3], não apenas porque carrega uma grande quantidade de informação acerca dos indivíduos e do ambiente, mas também por causa do baixo custo de implementação e baixo nível de intrusão dos sensores [1]. Ao mesmo tempo, o reconhecimento de sons domésticos é uma modalidade de reconhecimento de som ainda pouco estudada, se comparada ao reconhecimento de fala e de orador [3].

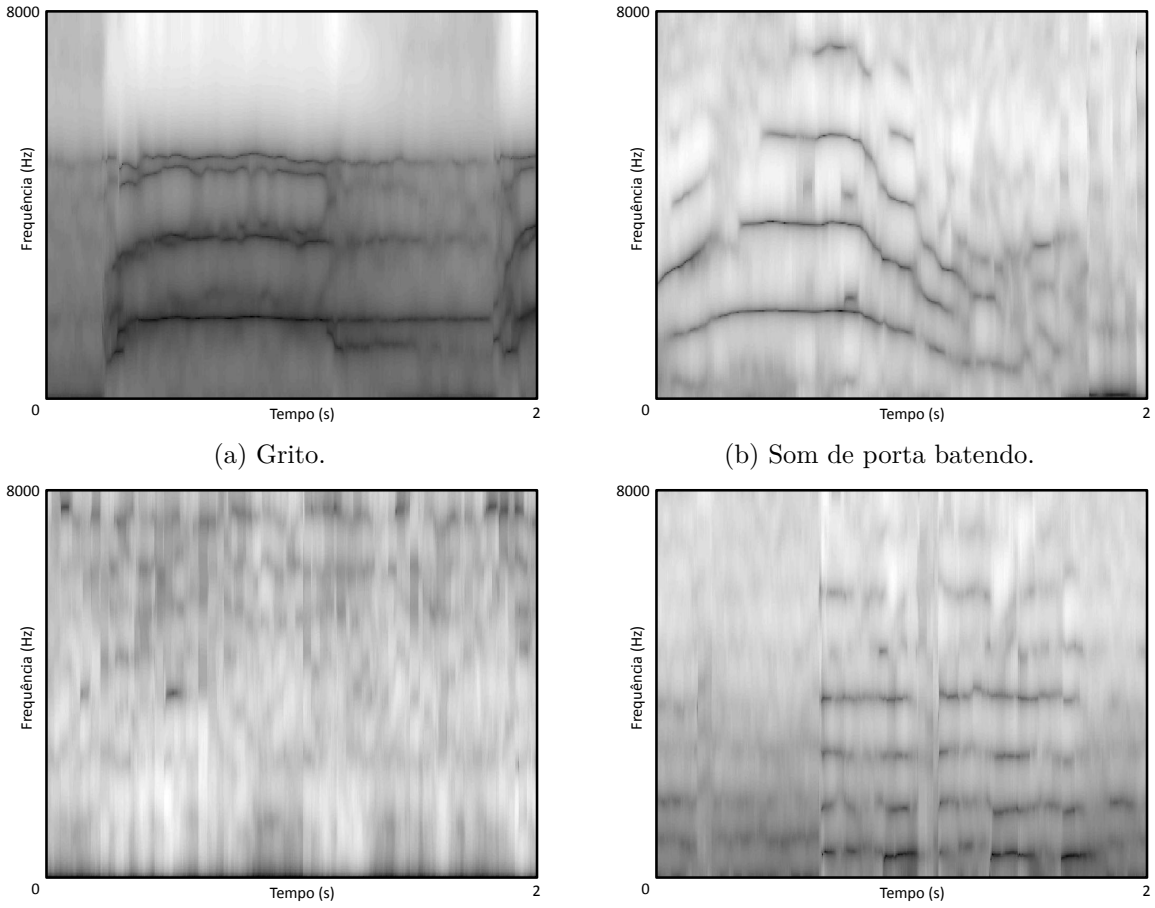
Vigilância sonora em ambiente aberto é uma modalidade análoga ao monitoramento

de sons domésticos, ambas são baseadas na verificação de sinais de áudio cotidianos. Essas modalidades não excluem sinais de voz, no entanto o interesse não é transcrever a fala nem verificar um indivíduo, mas sim rotular 'acontecimentos'. Por exemplo, um sistema público de vigilância de áudio pode receber como entrada sons de uma estação de trem, de tal forma que sua tarefa principal seria rotular trechos do sinal de acordo com um 'dicionário' predefinido de eventos acústicos, gerando um cenário como: "[pessoas conversando] [trem] [tiros] [gritos] [sirene]". Há diversas finalidades para sistemas de vigilância sonora, como segurança pública [15, 16], espionagem [17] e localização de fonte sonora [3].

Qualquer tipo de som pode entrar no dicionário de eventos acústicos (ou classes), no entanto é preciso levar em conta a alta variabilidade dos sinais de áudio (vide Figura 2). Por causa disso, a guia para elaboração do dicionário pode ser a percepção humana, isto é, sons facilmente identificáveis por seres humanos recebem rótulos cognitivos bem definidos, e criar uma mímica dessa rotulação pode ser uma boa meta mensurável para os sistemas automáticos. As bases de dados de sinais sonoros, utilizadas para testes de sistemas de reconhecimento de som, cumprem o papel de definir as classes de eventos acústicos. Tipos de sons que entram tipicamente em sistemas automáticos de vigilância de áudio são: tiros de armas de fogo, gritos, explosões, alarmes, batidas de porta, sons de carro, vidro quebrando, som de torneira d'água, entre outros.

No trabalho de Rabaoui et al. [18], por exemplo, foi desenvolvido um sistema de vigilância de áudio para reconhecer nove classes diferentes de sons (entre latidos, sons de crianças, telefone chamando etc) numa base de dados contendo um total de 1015 sinais, somando aproximadamente 30 minutos de áudio. Para gerar uma base grande, representativa e suficientemente diversa, os sinais foram retirados de três bibliotecas diferentes. Foram utilizados dois terços da base de dados para treinamento e um terço para teste, num sistema baseado em um classificador SVM de uma classe e uma combinação de vetores de características (MFCC, PLP, LPCC, wavelets, taxa de cruzamento por zero). Esse tipo de sistema é comumente utilizado em problemas de reconhecimento de fala ou, mais amplamente, em problemas de reconhecimento de padrões, isso quer dizer que apesar do monitoramento de áudio lidar com sinais que extrapolam os tradicionais sinais de fala e música, as técnicas tradicionais de aprendizado de máquina ainda funcionam razoavelmente bem para esses sinais exóticos.

Já no recente trabalho de Foggia et al. [12], um método menos tradicional de extração de característica é utilizado num sistema de vigilância de áudio. A proposta do artigo é um sistema que pode ser implementado mais facilmente em uma situação prática, isto é, que pode atuar de forma *online* numa série temporal fornecida por um microfone, por exemplo. Os autores testaram a proposta em três classes de eventos acústicos (grito, vidro quebrando e tiro de arma de fogo) e mais uma classe de ruído de fundo, de forma que o sistema verificava a todo momento qual classe estava ativa, para tanto, o esquema



(a) Grito.

(b) Som de porta batendo.

(c) Chaveiro balançando (choques entre chaves).

(d) Tosse humana.

Figura 2 – Espectrogramas de sinais corriqueiros em sistemas automáticos de monitoramento e televigilância de áudio.

de extração de característica computa um vetor que é uma mistura de vários descritores espectrais e energéticos. Esse vetor é calculado em *frames* de 32 ms do sinal, sobrepostos a uma taxa de 25%. O espaço dos vetores é quantizado utilizando agrupamento via *k-means*, cada centro encontrado é definido como a menor unidade fonética do sinal e chamado de *Palavra Auricular* (ou *Aural Word*). Um histograma das palavras auriculares é calculado numa janela de m segundos do sinal e utilizado como vetor de características num classificador SVM.

No que diz respeito à classificação de eventos sonoros, os sistemas atuais ainda estão longe de estabelecer um *framework* comum, como é o caso do reconhecimento de fala/orador (i.e., classificadores baseados em HMM, GMM, DTW, SVM ou redes neurais e extratores de características baseados em variantes da informação espectral) [3]. Talvez seja uma consequência (natural) desse estágio incipiente em que se encontra a arte, o fato das aplicações diretas das técnicas estado da arte advindas do reconhecimento de fala em problemas de monitoramento de áudio resultarem em taxas de classificação não compatíveis com as taxas dos sistemas de reconhecimento de fala, principalmente devido à alta variabilidade inerente dos sons ambiente [19].

1.2 A dinâmica espectral em sinais sonoros

A informação que um sinal sonoro natural carrega se concentra principalmente nas áreas em que suas características se alteram [5]. Em atividades relacionadas ao processamento de imagens, é notória a importância da etapa de detecção de bordas, por exemplo, onde a imagem original (no formato colorido ou monocromático) é transformada em uma imagem binária, indicando quais *pixels* compõem regiões de borda ou fronteira, assim como ilustrado na Figura 3. A detecção de borda pode ser vista como uma espécie de filtragem que destaca regiões de interesse na imagem, onde se concentram as importantes propriedades estruturais do cenário [20], etapa de processamento que parece ser fundamental no reconhecimento de formas pelo cérebro humano [21]. Nos problemas de aprendizado não supervisionado com imagens, onde as técnicas do *Deep Learning* se têm difundido bastante, as primeiras camadas de características extraídas da imagem são, geralmente, pequenos filtros de bordas localizados, indicando que áreas do sinal onde há mudanças acentuadas de suas características são áreas também consideradas de interesse [22, 23].

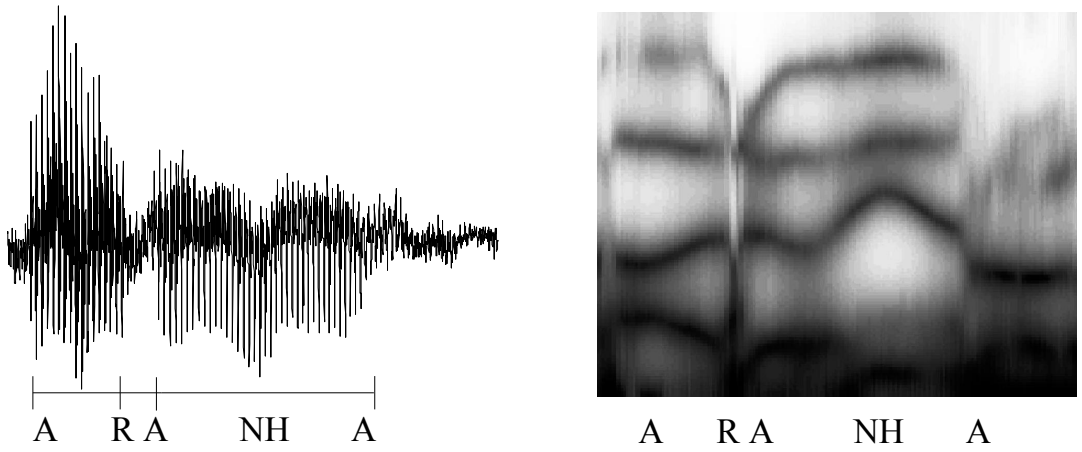
(a) Imagem original colorida (do disco "Paco de Lucia interpreta a Manuel de Falla").

(b) Imagem binária processada com o algoritmo Canny de detecção de bordas.

Figura 3 – Um exemplo da quantidade de informação que áreas de mudanças abruptas do sinal carregam. Na imagem 3b é possível reconhecer o texto e o formato dos objetos apenas sinalizando pequenas áreas que representam bordas.

A relevância das bordas em sinais de imagens encontra boa analogia no contexto de sinais sonoros, no qual as mudanças das características do sinal devem ser notadas ao longo do tempo. Essa analogia se torna ainda mais evidente através da representação visual que o espectrograma fornece. Ao contrário da representação visual tradicional de um sinal sonoro (gráfico da pressão do ar ao longo do tempo, como na Figura 4a), que fornece apenas uma informação vaga da intensidade e eventuais padrões repetitivos do sinal, a olho nu; a representação a partir do espectrograma permite, a um observador treinado

(como um fonoaudiólogo), a inferência visual de inúmeras características do som, desde o reconhecimento de fonemas até a detecção de algumas patologias da fala [24, 25].



(a) Gráfico de representação de um sinal unidimensional ao longo do tempo. Trata-se de um sinal de voz da elocução da palavra "aranha".

(b) O espectrograma é uma representação de um sinal no plano tempo-frequência (direção horizontal e vertical, respectivamente), a intensidade num ponto é representada pela intensidade em nível de cinza.

Figura 4 – A diferença da quantidade de informação visual que cada tipo de representação de um sinal sonoro carrega. Sabendo que esse sinal é uma elocução da palavra "aranha", na representação tradicional, só é possível localizar uma área de mudança rápida das características do sinal e podemos atribuir isso à elocução da consoante 'R'. Já na representação correspondente ao espectrograma, se percebem diversas nuances, como a presença de formantes e de suas modulações.

O exemplo mais notável de que a maior parte da informação está contida nas regiões de mudança em sinais de fala é, provavelmente, o efeito das consoantes. Consoantes são fonemas articulados com algum tipo de obstrução no trato vocal, como o som [t], pronunciado com uma obstrução feita com a ponta da língua, ou o som [f] e [s], pronunciados ao forçar o ar num canal estreito. Diferente das vogais, que podem ser pronunciadas isoladamente, as consoantes, de forma geral, dependem de um fonema vocálico para serem pronunciadas. Esta afirmação pode soar um tanto inusitada, afinal se em uma elocução de uma palavra (como 'aranha' da Figura 4) estão presentes fonemas consonantais e vocálicos, deve ser possível localizar o instante em que cada um é executado unicamente, e daí reproduzí-los ou extrair suas características. Apesar disso ser completamente verdadeiro para fonemas vocálicos, de onde surgem os conceitos de formantes, para consoantes, em contraste, nem sempre é possível localizar um instante preciso em que elas são pronunciadas. O conceito de fonema (menor unidade sonora da fala) leva a crer que a construção de uma palavra falada é análoga à construção de uma palavra escrita, como se para falar a palavra 'aranha' bastasse pronunciar, isolada e sequencialmente, os fonemas [a] + [r] + [a] + [nh] + [a], quando na verdade fonemas consonantais são articulados de formas diferentes, a depender de sua vizinhança (transição das formantes), ou não soam isoladamente de forma

familiar (efeito de energia distribuída por todo o espectro, *burst spectrum*) [26]. Apesar da grande quantidade de informação encontrada na literatura a respeito das características das consoantes, ainda existe ambiguidade e contradição em sua descrição [27, p. 834].

Um exemplo notório de como as consoantes desempenham um papel fundamental ao transmitir informação, tanto na forma oral quanto na forma escrita, é o experimento de supressão de vogais e consoantes num texto, ilustrado na Figura 5. É pouco provável que um indivíduo consiga ler um texto sem vogais ou sem consoantes, no entanto o papel das consoantes no entendimento do conteúdo é crítico, ao contrário das vogais. No exemplo da Figura 5, é possível compreender algumas palavras no texto, mesmo com a omissão das vogais. Já no texto sem consoantes, a assimilação é quase nula. Vejamos, por exemplo, a palavra 'boca', que resulta em 'bc', com supressão das vogais, e 'oa', com supressão das consoantes. A quantidade de palavras que podem ser construídas com o padrão 'oa' (e.g. porta, botar, boca, proa, forma etc) é muito maior que a quantidade de palavras que podem ser construídas com o padrão 'bc' [2], além disso, quanto mais palavras são reconhecidas corretamente num texto escrito em linguagem natural, mais fácil se torna reconhecer outras palavras no mesmo texto, devido à interpretação de contexto (ou redundância). Um experimento oral semelhante pode ser encontrado na brincadeira infantil da cantiga de roda 'O Sapo Não Lava o Pé', em que o verso é cantado usando apenas uma vogal de cada vez (e.g., 'A sapa naa lava a pá, naa lava parca naa car'), isto é, descartando a informação que as vogais carregam na frase.

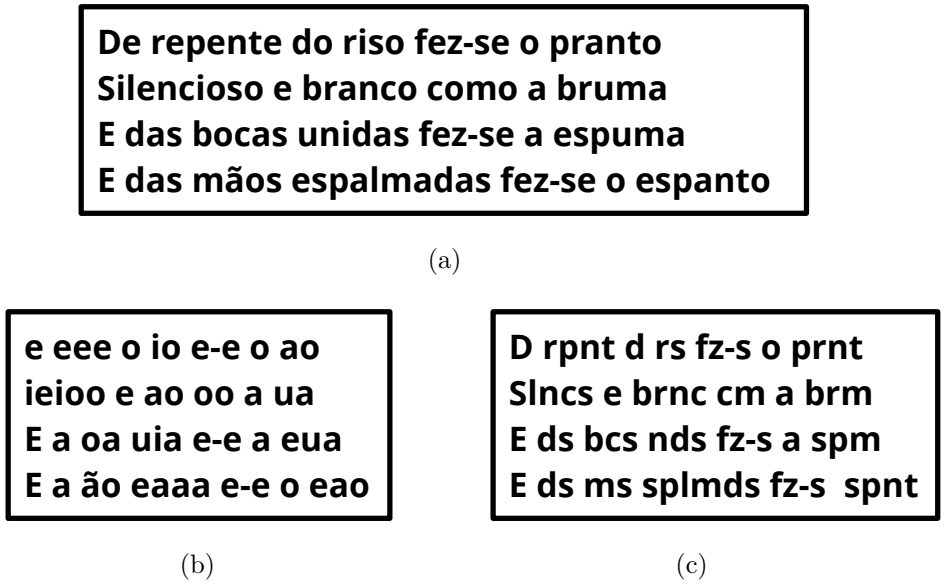


Figura 5 – Trecho do Soneto da Separação de Vinícius de Moraes e sua reprodução com supressão de consoantes e vogais (Figura 5b e 5c, respectivamente). O texto com supressão de vogais ainda conta com algumas vogais para não suprimir palavras de uma letra só. Um exemplo de que consoantes transmitem mais informações que vogais em texto escrito.

[2] A título de exemplo, uma pesquisa exaustiva em uma lista de 308029 palavras em português, disponível em <https://dl.dropboxusercontent.com/u/42709342/pt_BR.dic>, resulta em 914 palavras com o padrão 'oa' de vogal e apenas 68 com o padrão 'bc' de consoantes.

Tendo em vista as características dos fonemas consonantais aqui discutidas, deve ser compreensível a dificuldade de reconhecer consoantes automaticamente. Comparada à forma como o ser humano reconhece consoantes, a máquina já começa em desvantagem, uma vez que a percepção da fala é potencializada por estímulos visuais [3] e por interpretação de contexto. Ainda assim, sistemas atuais de reconhecimento de consoantes chegam a uma taxa de acerto na faixa de 80% [29].

Toda essa análise a respeito das consoantes pode parecer desnecessária, visto que o propósito deste trabalho não está focado em sinais de voz, nem em reconhecimento de fala. No entanto, a área de reconhecimento de sons domésticos é razoavelmente nova e técnicas exclusivas dessa área ainda estão em desenvolvimento, isto é, trabalhos na área de reconhecimento de sons domésticos utilizam técnicas tradicionais de reconhecimento de fala e de orador (MFCC, HMM, SVM, GMM, quantização vetorial etc) como patamar inicial de pesquisa [1,30,31]. Além disso, a importância de se estudar as consoantes se deve ao fato de que a cognição humana das variações espectrais é refinada pela necessidade de percepção das consoantes, que carregam a maior parte da informação de fala. Logo, não é insensato assumir que o humano usa esse mesmo refinamento cognitivo no processamento robusto de outros sons.

Finalmente, é de se esperar que a informação que as consoantes carregam estejam codificadas em inúmeros padrões de 'mudança' no espectro sonoro, em outras palavras, a dinâmica espectral. Apesar disso, durante muito tempo, têm sido adotadas técnicas de extração de características de sinais de som baseadas no envelope de potência do espectro como principal portador de informação linguística em sistemas automáticos de reconhecimento de fala (como o LPC, MFCC, PLP etc) [32,33]. Técnicas baseadas em suavização do espectro descartam muitas informações importantes do espectro, como estruturas não estacionárias, transientes, *pitch* [34]. Por consequência, a relação entre características qualitativas do som e o seu envelope espectral de curto termo não é direta, como se pode imaginar. Como resultado, tais técnicas demandam mecanismos complexos de mineração de dados e aprendizado de máquina, que são prevalecentes em sistemas estado da arte. Apesar de grandes avanços, sistemas automáticos de reconhecimento de fala e orador ainda são sensíveis à reverberação, à distorção do canal, a ruídos aditivos e a particularidades da mensagem ou do orador — problemas com os quais o ouvido humano consegue lidar razoavelmente bem [5].

Ainda no âmbito de sinais de fala, sabe-se que, *grosso modo*, variações lentas de energia no sinal de voz estão associadas aos movimentos de articulação das sílabas [35,36]. Partindo desse pressuposto, Dudley [4] desenvolveu um método simples para sintetizar voz humana, estudo que resultou na criação do primeiro sintetizador de voz eletrônico, o

[3] Estímulos visuais podem também atrapalhar o reconhecimento de fala pelo ser humano, como é o caso do efeito McGurk [28]

VODER. Para captar a informação fonética do som, não basta capturar o contorno de potência do sinal, pois assim a informação espectral seria perdida. A inovação de Dudley foi captar o contorno de potência do sinal na saída de uma série de filtros passa-faixa, dividindo o espectro em um número discreto de canais. Com esta técnica foi possível analisar o sinal de voz de forma mais concreta (e.g., para criar o som da vogal [u] basta ativar canais específicos a amplitudes específicas), possibilitando a 'recriação da fala' [4] e, por consequência, a síntese de voz. Dudley foi capaz de demonstrar que a informação essencial da fala está encapsulada em padrões de modulação na escala de 40 milissegundos, distribuídos em apenas 10 canais espectrais [36]. A Figura 6 ilustra um espectrograma reconstruído a partir dos padrões de modulação por banda, de acordo com Dudley.

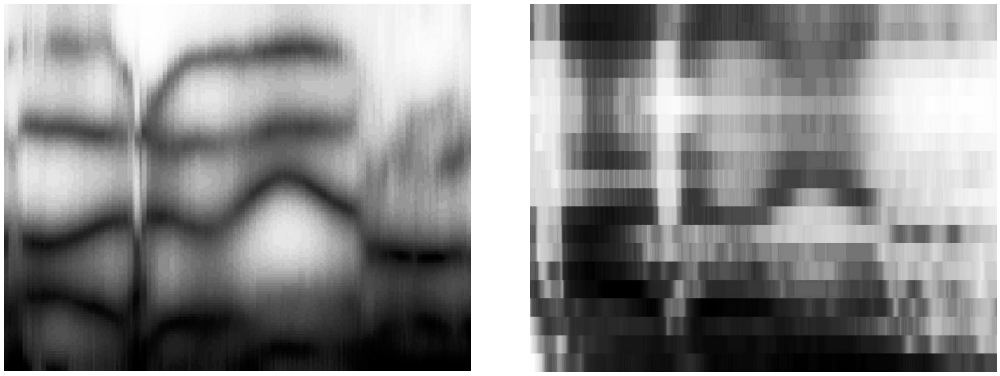


Figura 6 – Espectrograma da elocução da palavra 'aranha' à esquerda e, à direita, o espectrograma construído a partir dos padrões de modulação de Dudley (aqui foram usados 20 canais para uma melhor apresentação da figura, embora apenas 10 canais sejam suficientes, segundo Dudley). Ambas figuras representam frequências, na faixa de 0 a 4000 Hz, na direção vertical e o tempo no eixo horizontal.

Os padrões de modulação de Dudley dão ênfase a um tipo de informação complementar às características baseadas em envelope de potência do espectro (como o MFCC), no sentido que um enfatiza a variação de energia ao longo do tempo, enquanto o outro enfatiza a variação de energia em função da frequência. Apesar disso, cada um dos métodos pode ser adaptado de forma a se aproximar do outro [4]. Por exemplo, os padrões de modulação de Dudley podem ser usados para dividir o espectro em um grande número de canais estreitos, o que implicaria numa maior resolução da distribuição de energia pelo espectro, ao preço de uma perda de resolução em tempo. Em contrapartida, as características baseadas em envelope de potência do espectro podem ser extraídas em uma sequência temporal de janelas sobrepostas, produzindo assim uma representação da variação da energia por banda ao longo do tempo, ao preço de uma perda da resolução em frequência. No entanto, diferentemente dos métodos baseados em envelope de potência do espectro, a informação dos padrões de modulação de energia por banda não têm uma

[4] Na verdade, qualquer método utilizado para descrever sinais de som deve, de alguma forma, englobar o caráter informativo da energia em função da frequência e em função do tempo, se o intuito for reconhecer automaticamente sinais de som.

codificação bem estabelecida na literatura, e essa codificação é relevante para uso na etapa de extração de características [36]. Na verdade, usualmente, quando engenheiros de sistemas de reconhecimento automático de som desejam incluir a informação da dinâmica espectral no processo, não são os extratores de características que mudam, ao invés disso, técnicas que analisam o caráter evolutivo dos vetores de características são incluídas (como o HMM [37] e os coeficientes de regressão linear [38]).

As tradicionais técnicas de extração de características de sinais de som baseadas no envelope de potência do espectro [39] têm uma característica em comum, além da suavização do espectro: todas elas descartam a informação de fase do sinal (informação que vem pareada com a informação de potência harmônica na transformada de Fourier). Isso talvez se deva à controversa — e pouco conhecida — Lei Acústica de Ohm. Ela afirma que dois sons com o mesmo padrão de amplitude em seus harmônicos (ou o módulo da representação pela Transformada de Fourier), mas com padrões diferentes na representação de fase da Transformada de Fourier, soarão identicamente para um ouvinte, apesar das formas de onda serem completamente diferentes [40, p. 114]. Na verdade, o que foi afirmado por Georg Ohm em 1843, simplificadamente, é que o ouvido humano é sensível à frequência e à amplitude das ondas acústicas e, portanto, é capaz de perceber diferenças em sons com frequências e amplitudes diferentes [41]. Uma afirmação que parece óbvia, dados os alicerces de conhecimento atual; no entanto, para o século XIX (período em que os conceitos da Transformada de Fourier ainda eram novidade), essa era uma observação notável. Helmholtz [42] foi o responsável por adicionar as indagações a respeito da insensibilidade auditiva à fase na Lei Acústica de Ohm [43, 44], principalmente porque os importantes experimentos desenvolvidos por ele analisavam sons 'musicais' em estado estacionário.

Para testar a sensibilidade do ouvido humano à fase, o experimento possivelmente mais simples (similar ao desenvolvido por Helmholtz) seria aquele em que dois sinais são sintetizados a partir de uma soma de senoides de mesma frequência e amplitude, mas com fases diferentes [45]. Os sinais, ao serem apresentados sequencialmente (com um intervalo de silêncio entre ambos), não se distinguem de forma clara ao ouvinte, o que seria evidentemente o caso, para uma mudança de frequência [40]. Um sinal constituído com apenas dois senos e com uma variação abrupta de 180º na frequência mais alta está ilustrado na Figura 7. É verdade que esse experimento pode facilmente levar à conclusão da insensibilidade à fase [5], mas há um experimento igualmente simples que refuta essa afirmação: a inversão temporal do sinal. Qualquer sinal revertido no tempo tem a mesma distribuição energética no espectro que o sinal original, mas o gráfico de fase invertido.

[5] Mudar a relação de fase entre os harmônicos de um som periódico pode alterar o timbre [46]; no entanto, esse efeito é muito sutil, e é geralmente inaudível em ambientes reverberantes normais em que a relação de fase é atenuada. Uma explicação comum para a insensibilidade à fase experimentada por Helmholtz seria o alto intervalo de tempo entre a apresentação dos dois sinais, atenuando ainda mais a percepção do efeito da fase [47].

Logo, não é difícil imaginar que praticamente qualquer som complexo (isto é, um sinal não estacionário) soará diferente da sua versão revertida temporalmente [48]. A questão da sensibilidade à fase ainda é um debate em aberto (apesar da constatação de vários efeitos relacionados, como a dependência do envelope de onda na percepção do som [46]), mas a conclusão parcial à qual chegamos aqui é que, fora casos extremos, a variação de fase não provoca efeitos relevantes na percepção do som (segundo Plomp e Steeneken [49], "o maior efeito que a fase produz no timbre é menor que o efeito de inclinação de 2 dB por oitava no padrão de amplitude"), e talvez por causa disso, historicamente, essa informação tem sido descartada.

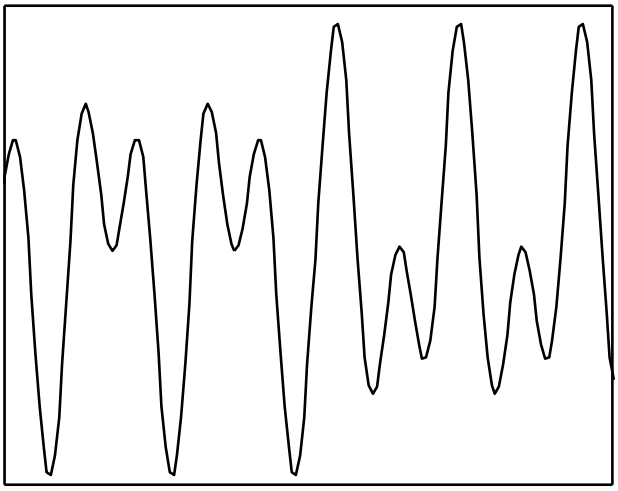


Figura 7 – Um sinal sintetizado a partir de um seno a 250 Hz e seu primeiro harmônico. Após um intervalo de tempo, o harmônico muda em 180º sua fase, gerando uma região abrupta de transição (que pode ser notada como uma consoante). Escutando em um fone de ouvido, a maioria das pessoas percebe a mudança de fase.

Outra característica importante dos sinais de som é a variabilidade dos períodos de estado estacionário, tanto em sinais de voz como em sinais de música e sons domésticos [50, 51]. Isso reflete diretamente no entendimento do atributo sonoro do timbre, fundamental para qualquer sistema de reconhecimento de som. O timbre está intimamente ligado aos períodos de estado estacionário do som, e caracterizá-lo apenas por este ponto de vista já seria uma tarefa difícil (consistindo basicamente em encontrar todos os possíveis formatos espectrais distintos, do ponto de vista psicoacústico) [52, 53] . No entanto é notória a importância dos períodos transitórios na percepção do timbre (mais especificamente o ataque e o decaimento). Devido à natureza interativa da geração de sons (pois todas as conexões mecânicas de objetos vibrantes serão incorporadas ao som resultante), as propriedades que podem caracterizá-los são inúmeras, é pouco provável que uma propriedade — ou a combinação de algumas poucas — possa determinar por si só a natureza do timbre [54]. Apesar disso, a maioria dos estudos de análise de timbre (desde as primeiras pesquisas de Helmholtz [42]) têm se limitado a examinar sinais em estado estacionário, negligenciando os tipos de som com mudanças dinâmicas ao longo do tempo,

chamados de sinais naturais [51].

Sem dúvida, o aspecto da dinâmica espectral é decisivo na percepção de sons, no entanto ele não é posto comumente em xeque quando se analisam sistemas de reconhecimento de sons [3, 12, 18], apesar de várias ferramentas tradicionais de reconhecimento de sons incorporarem aspectos de dinâmica espectral (como o HMM, os coeficientes delta, os DWTC, o *Relative Spectral (RASTA), Perceptual Linear Prediction*, entre outros). Ainda assim, há vários trabalhos focados no entendimento da dinâmica espectral, como é o caso dos trabalhos de Hynek Hermansky [5], de Sadaoki Furui [38, 50] e também os trabalhos clássicos de Homer Dudley [4] e Harvey Fletcher [55], entre outros.

1.3 Organização do trabalho

Este trabalho de mestrado está divido em três capítulos, além da introdução. No primeiro capítulo serão discutidos alguns conceitos teóricos da tarefa de reconhecimento automático de sons, assim como algumas técnicas e métodos que foram utilizados durante os experimentos de pesquisa realizados ao longo do mestrado. Ainda mais, será explicado o trabalho usado como referência inicial desta pesquisa.

No segundo capítulo serão discutidos todos os experimentos realizados ao longo do trabalho, expondo os resultados obtidos e analisando-os. Ao todo, foram quatro experimentos mais relevantes: a reprodução do trabalho de Sehili et al. [1], a análise dos coeficientes de regressão linear na classificação de sons domésticos, a análise do HMM na classificação de sons domésticos e o desenvolvimento do extrator de características chamado PEEB.

No quarto e último capítulo, serão postas as conclusões e as perspectivas de pesquisa futura com tema relacionado ao deste trabalho.

2 ASPECTOS GERAIS DO RECONHECIMENTO AUTOMÁTICO DE SONS E O PATAMAR INICIAL DA PESQUISA

Ao desenvolver um sistema de reconhecimento de sons, os engenheiros têm de lidar com uma série de desafios que, em sua maioria, são naturais dos problemas de aprendizado de máquina. No entanto, os sinais de áudio acrescentam particularidades aos desafios que estão associados principalmente à natureza aleatória de alta variabilidade e contínua do som. Ruídos estão sempre presentes em sinais de áudio coletados naturalmente, seja pelos sons de fundo e da reverberação, seja pelas distorções do microfones e do canal, ou até pela conversão A/D nos casos mais comuns hoje em dia, em que os sinais são convertidos para o formato digital. Além do mais, a codificação de informação em sinais de som (como no processo da fala, por exemplo) gera sinais que mudam rapidamente suas características, o que induz um tipo de análise em que a estacionariedade só é válida, aproximadamente, em intervalos de tempo na ordem de milésimos de segundo. Por causa disso é difícil determinar quais parâmetros são relevantes num sinal de som. Por exemplo, consideremos o caso de uma pessoa utilizando um assistente pessoal e falando ao microfone do seu celular “Como estará o tempo amanhã?”. Se a captação der certo, o programa terá de lidar agora com uma sequência de números que representa, aproximadamente, 2 segundos de áudio, algo por volta de 16000 medidas escalares representadas digitalmente, se o sinal for amostrado a uma taxa de 8 kHz. Mas onde e como está codificada a informação útil da mensagem, no meio desses 16000 números? Essa é uma das questões fundamentais no estudo do reconhecimento de sons, e apesar dela estar razoavelmente bem explicada (afinal de contas, os assistentes pessoais estão funcionando), ainda há espaço para muita evolução nesse processo. Na prática, essa questão é inicialmente endereçada por meio de duas técnicas: a extração de características e a segmentação de áudio.

Extração de características é uma etapa que envolve a redução da quantidade de recursos necessários para descrever um grande conjunto de dados. Ao realizar a análise de dados complexos, um dos principais problemas é o número de variáveis envolvidas. Análise com um grande número de variáveis geralmente requer uma grande quantidade de memória e poder de processamento por parte do classificador, sob pena deste não gerar uma boa generalização do modelo. Extração de características é um termo geral que engloba os métodos que contornam os problemas supracitados, mas que ainda assim conseguem descrever os dados de forma satisfatória ao propósito da classificação. Para o caso de sinais de áudio, a etapa de extração de características é praticamente imprescindível, num

sistema de reconhecimento, pois a maior parte da informação está contida nos padrões de distribuição energética espectro-temporal do som [4]. Por causa disso, os principais extratores de características para sinais de áudio procuram encontrar uma representação de baixa dimensão do espectro do sinal. Uma rotina comum dentre tais métodos segue os seguintes passos (também ilustrados na Figura 8):

1. O sinal bruto é pré-processado (com um filtro de ênfase, por exemplo) e dividido em janelas sobrepostas, de duração na faixa de dezenas de milissegundos;
2. Em cada janela é calculado um vetor de baixa dimensão que representa o espectro do sinal de forma suavizada;
3. O sinal bruto é finalmente representado por uma matriz de tamanho $D \times N$, isto é, N vetores de dimensão D concatenados.

MFCC, LPC, LPCC, RASTA, PLP são alguns dos extratores de características de sinais de áudio que incluem esses procedimentos comuns [3]. Neste trabalho, o extrator de características utilizado na maioria dos testes é o MFCC [56]. Já os espectrogramas que ilustram este trabalho (como as figuras 4b e 2) foram calculados utilizando um extrator de características no formato do LPC. A representação suavizada do espectro no LPC é fornecida por p coeficientes, ajustados de forma que a combinação linear de p amostras do sinal possa prever a próxima amostra com um certo nível de tolerância, como na equação

$$a_1 x(n-1) + a_2 x(n-2) + ... + a_p x(n-p) = x(n) + \epsilon.$$

Uma vez ajustados, os coeficientes a_i descrevem um sistema discreto de ordem p cujo retrato espectral representa uma versão suavizada do espectro do sinal de entrada. Para gerar o espectrograma baseado em LPC, basta criar uma matriz em que cada coluna contenha um vetor que representa as amostras do espectro suavizado de cada janela do sinal, e que cada elemento da matriz represente um *pixel* de uma imagem em escala de cinza.

Vale lembrar que quando o sinal de som é representado por uma matriz de características (ou por qualquer outro formato de um extrator de características), uma grande quantidade de informação tem de ser descartada, afinal esta é a proposta da etapa de extração de características. A depender da finalidade, cada processo de extração de características pode focar num atributo diferente do sinal de som. A característica que o MFCC foca, por exemplo, é o contorno espectral de potência, como citado anteriormente. Por causa da sua construção e do seu desempenho, o MFCC é um extrator de características bastante usado em sistemas e na literatura, para diversas finalidades [3, 57]. Levando em conta a dinâmica espectral, apenas um vetor de coeficientes do tipo MFCC não é capaz de representar com clareza a evolução temporal do espectro (pois sua análise é feita em

janelas do sinal), além de descartar informações potencialmente relevantes do espectro (como é o caso do *pitch*). Mesmo assim, muitas informações de dinâmica espectral ainda são capturadas pela matriz de características do MFCC, principalmente quando se usa um classificador que, de alguma forma explícita ou não, inclui a modelagem da evolução dinâmica desses vetores de coeficientes.

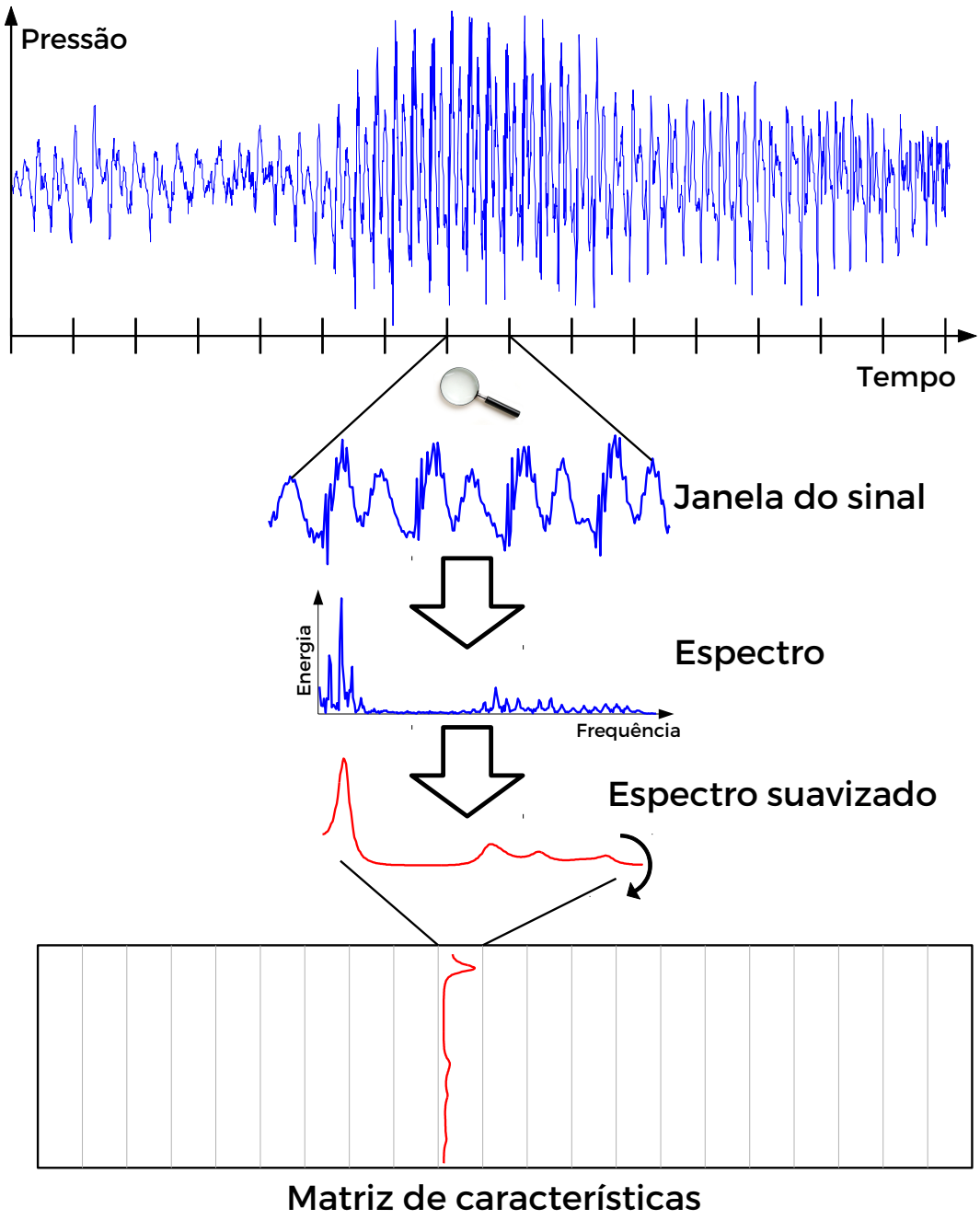


Figura 8 – Rotina de extração de características baseada em suavização do espectro de um sinal de áudio.

O processo de extração de características tenta evidenciar como está codificada a informação no sinal, já o processo de segmentação de áudio tenta evidenciar onde está a informação no sinal, geralmente ao longo do tempo. Voltando para o exemplo do assistente

pessoal, o sinal de áudio que contêm a elocução da frase "Como estará o tempo amanhã?" deve, provavelmente, iniciar com um ruído de fundo antes da voz, ter alguns instantes de pausa durante a fala e terminar com o ruído de fundo novamente. Se o objetivo do assistente pessoal é reconhecer o que se fala, os trechos ruidosos sem elocução não devem fazer parte da análise, portanto eles devem ser descartados na etapa de segmentação de áudio.

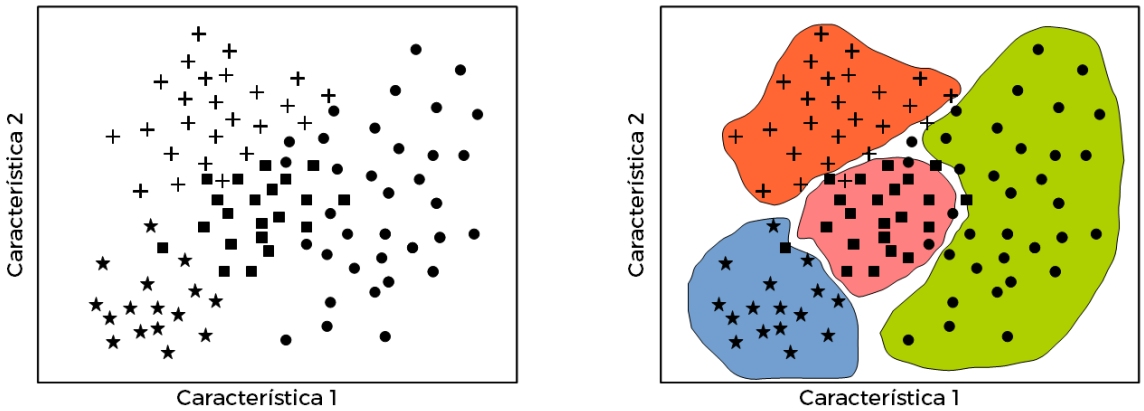
As técnicas mais simples de segmentação usam o perfil de energia do sinal, ou volume, e as taxas de cruzamento por zero ao longo do tempo, para determinar as regiões de interesse. Técnicas mais sofisticadas podem utilizar vetores de características diversos para determinar áreas de interesse, seja em sinais de fala, de música ou de vigilância de áudio [58–60]. Outra possibilidade de segmentação é utilizar uma classe que englobe o ruído de fundo, de modo que na etapa de classificação, as áreas sem conteúdo para o sistema sejam rotuladas com um único rótulo atribuído a todos os sinais que não sejam de interesse [12, 61]. Neste trabalho, a etapa de segmentação de áudio não é analisada explicitamente por não ser o foco do estudo, e por causa do formato da base de dados utilizada para os testes (mais explicações sobre a base de dados são fornecidas na seção 2.2).

2.1 Classificadores baseados em Mistura de Gaussianas e Modelo Oculto de Markov

Em Aprendizado de Máquina e Reconhecimento de Padrões, classificação é o problema de identificar a que categoria uma nova observação pertence, a partir de um conjunto de dados de treinamento que contem observações cujas categorias são conhecidas. A etapa de ajuste de um classificador a partir desses dados é considerada como parte do domínio de Aprendizado Supervisionado [62, 63], pois usa padrões já rotulados para inferir a classe de um novo padrão não rotulado.

O leque de opções de algoritmos de classificação usados atualmente é extenso, desde os classificadores lineares, máquinas de vetores suporte, métodos baseados em *kernel*, até as redes neurais, entre outros. Apesar de tantas opções de métodos de classificação, a etapa de classificação pode ser resumida como uma tarefa de particionamento do espaço vetorial de características gerado pelas observações (ou padrões), no intuito de determinar regiões do espaço para cada classe. Num problema de monitoramento de sons domésticos, por exemplo, se um vetor de duas dimensões pudesse representar todas as diferenças entre quatros classes de sons (gritos, vidro quebrando, porta batendo e água escorrendo), como na Figura 9, a tarefa do classificador seria dividir o espaço de características em quatro regiões disjuntas, uma para cada classe, de forma que se minimize o erro de classificação a partir dos padrões rotulados. Além disso, o classificador deve determinar as regiões

também levando em conta a capacidade de generalização do sistema. Na Figura 9b, por exemplo, há um padrão do formato 'quadrado' dentro da região dos padrões formato 'estrela', e um classificador poderia englobar esse padrão para torná-lo parte da região rosa, diminuindo o erro de classificação dos padrões de treinamento, no entanto isso criaria um sistema sem uma boa generalização, pois geraria regiões entrelaçadas que provavelmente só melhorariam a taxa de classificações corretas para os dados de treinamento.



(a) Espaço de características ilustrativo de duas dimensões. Cada padrão é representado por um ponto no gráfico, os diferentes formatos representam classes diferentes (círculo, quadrado, cruz e estrela).

(b) A tarefa de um classificador é determinar regiões no espaço de características que representem cada classe. Nesta figura, as classes 'círculo', 'estrela', 'quadrado' e 'cruz' são delimitadas pelas regiões de cor verde, azul, rosa e laranja, respectivamente.

Figura 9 – A etapa de classificação.

Devido à natureza probabilística e ruidosa dos sinais e dos equipamentos de aquisição, muitos classificadores têm uma base teórica estatística, eles são chamados de classificadores probabilísticos. Dentre esses, o classificador bayesiano é aquele que minimiza a probabilidade de erro de classificação [64]. Os classificadores projetados a partir desse conceito determinam a categoria de um padrão desconhecido como sendo aquela mais provável [63]. Em linguagem matemática, um problema com L classes $C_1, C_2, ..., C_L$, a probabilidade de um padrão $\mathbf{x}$ pertencer a uma classe C_i é chamada de probabilidade *a posteriori*, e pode ser escrita na forma de probabilidade condicional $P(C_i|\mathbf{x})$. O Teorema de Bayes permite escrever esta probabilidade como

$$P(C_i|\mathbf{x}) = \frac{P(C_i)p(\mathbf{x}|C_i)}{p(\mathbf{x})}$$

em que $P(C_i)$ é a probabilidade *a priori*. No caso da base de treinamento (amostra) ter sido amostrada criteriosamente para guardar as proporções da população de padrões amostrada, $P(C_i)$ pode ser estimada a partir do número total de padrões de treinamento N e o número de padrões da classe N_i por meio de $P(C_i) \approx \frac{N_i}{N}$, $p(\mathbf{x})$ é a função densidade de probabilidade (PDF) de $\mathbf{x}$ e $p(\mathbf{x}|C_i)$ é a PDF condicionada à classe C_i [62].

Como os classificadores estatísticos determinam a categoria de um novo padrão encontrando a maior probabilidade *a posteriori* $P(C_i|\mathbf{x})$, utilizando o Teorema de Bayes

é possível converter o problema de encontrar $P(C_i|\mathbf{x})$ no problema dual de encontrar a probabilidade *a priori* de cada classe e a PDF $p(\mathbf{x}|C_i)$. Dessa forma, para fins de classificação dos padrões observados, devem ser avaliadas as desigualdades

$$P(C_i)p(\mathbf{x}|C_i) \lessgtr P(C_j)p(\mathbf{x}|C_j) \tag{2.1}$$

Num problema envolvendo classes equiprováveis, os *a priori* são todos iguais e, portanto, podem ser descartados da inequação 2.1. Logo, para projetar um classificador bayesiano, a etapa decisiva é estimar a PDF $p(\mathbf{x}|C_i)$, que é a distribuição dos dados de uma certa classe C_i. O modelo probabilístico de mistura de gaussianas (GMM) é um modo paramétrico de representar PDFs de vários formatos, que quando usado em conjunto com o algoritmo EM torna-se um estimador de PDF bastante flexível, pois quase qualquer função densidade de probabilidade contínua pode ser aproximada por uma GMM a uma precisão arbitrária [62]. Um GMM $\theta = \{\mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ formada por M componentes Normais de dimensão D é descrito por $M(1 + D + D^2)$ parâmetros: cada componente Normal tem um vetor de média $\boldsymbol{\mu}_m$, uma matriz de covariância $\boldsymbol{\Sigma}_m$ e um coeficientes de peso c_m. A PDF definida pelo GMM é escrita como uma combinação linear de todas as componentes[1]

$$p(\mathbf{x}) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

Um dos algoritmos mais usados para inferir os parâmetros de uma GMM é o EM. Um esboço deste algoritmo, no contexto do GMM, será descrito a seguir; mais detalhes sobre o EM são fornecidos em [62, 63].

Dados N vetores de treinamento de dimensão D, e uma GMM com M componentes, para maximizar a verossimilhança dos parâmetros do modelo em relação aos dados de treinamento, é preciso:

1. Inicializar as médias, matrizes de covariância e pesos da GMM;
2. Calcular a matriz na forma

$$\gamma(m, n) = \frac{c_m \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{j=1}^{M} c_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

[1] Para $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$

3. Atualizar os parâmetros da GMM, seguindo

$$\widehat{\boldsymbol{\mu}}_m = \frac{1}{N_m} \sum_{n=1}^{N} \gamma(m,n) \cdot \mathbf{x}_n$$

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{1}{N_m} \sum_{n=1}^{N} \gamma(m,n) \cdot (\mathbf{x}_n - \widehat{\boldsymbol{\mu}}_m)(\mathbf{x}_n - \widehat{\boldsymbol{\mu}}_m)^T$$

$$\widehat{c}_m = \frac{N_m}{N}$$

em que

$$N_m = \sum_{n=1}^{N} \gamma(m,n)$$

4. Verificar a log-verossimilhança

$$\log \mathcal{L}(\theta|\mathbf{X}) = \log p(\mathbf{X}|\theta) = \sum_{n=1}^{N} \log \left\{ \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \right\},$$

e caso não atinja o critério de convergência, voltar ao passo 2.

Uma fragilidade do estimador de PDF baseado em GMM ajustado com EM se encontra na forma de inicialização dos parâmetros. Devido à quantidade de parâmetros a serem estimados e à complexidade do modelo, o algoritmo EM pode convergir para valores de máxima verossimilhança locais no hiperespaço de parâmetros, ou seja, a qualidade do modelo estimado é altamente dependente da inicialização do algoritmo. Na prática, é comum a inicialização aleatória dos parâmetros do GMM adaptado através do EM (levando em conta a localização dos vetores de treinamento), sendo repetida diversas vezes, gerando vários conjuntos de parâmetros com diferentes verossimilhanças.

Uma vez estimada as PDFs $p(\mathbf{x}|C_i)$ de cada classe e assumindo classes equiprováveis, para classificar um padrão desconhecido, basta calcular as densidades de probabilidade $p(\mathbf{x}|C_i) = p(\mathbf{x}|\theta^i)$ para cada classe, e designar o padrão à classe de maior probabilidade *a posteriori*. O resultado dessa comparação reflete no particionamento do espaço vetorial de características, como dito anteriormente. As fronteiras de decisão são as regiões definidas pela igualdade $P(C_i)p(\mathbf{x}|C_i) = P(C_j)p(\mathbf{x}|C_j)$. Há uma sutileza, no entanto, quando a tarefa é classificar sinais sonoros, visto que a etapa de extração de características geralmente cria uma sequência de padrões (e não apenas um) que representam determinado sinal de áudio. É possível classificar este conjunto de padrões com o GMM — assumindo independência entre os padrões, ao longo do tempo — calculando $p(\mathbf{X}|\theta) = \prod_{t=1}^{T} p(\mathbf{x}_t|\theta)$. Dessa forma, a informação de evolução temporal dos padrões é praticamente perdida, pois, do ponto de vista da GMM, não importa a ordem em que os padrões são apresentados, eles serão todos misturados no produtório, evidentemente, sem importar a ordem dos fatores.

O classificador bayesiano baseado em GMM é bastante utilizado em sistemas de reconhecimento de som [1, 3, 19] pois gera resultados satisfatórios em várias aplicações e

também porque vários pacotes computacionais têm o GMM já implementado, facilitando o seu uso tanto para os profissionais quanto para os amadores na área.

Quando se deseja incorporar a informação da evolução temporal da matriz de característica de um sinal de áudio, uma alternativa é a utilização do HMM. O HMM é um tipo de modelagem estocástica que se encaixa muito bem em sinais não estacionários. Consiste numa rede (ou grafo orientado) onde os nós são chamados de estados e as arestas representam transições entre os estados, como na Figura 10. Diferente da cadeia de Markov, os estados não são aparentes no HMM, mas sim as observações, que são apresentadas no modelo a partir de uma probabilidade de emissão de observação em cada estado. Por exemplo, um dado HMM composto por N estados, descritos pela matriz de transição $A = \{a_{ij}\}$, pela distribuição de probabilidade de estado inicial $\boldsymbol{\pi} = \{\pi_i\}$ e pelas distribuições de probabilidade de observação por estado $B = \{b_i(\mathbf{x})\}$, pode gerar uma sequência de observações $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$. Para gerar sua primeira observação, instancia-se o estado inicial $q_1 = S_k$ via $\boldsymbol{\pi}$ que, por sua vez, instancia a primeira observação $\mathbf{x}_1$ via $b_k(\mathbf{x})$. As próximas observações são geradas de forma semelhante, no entanto os próximos estados são determinados pela distribuição de transição do estado atual, no caso, a linha k da matriz A, para o estado $q_1 = S_k$ [2]. As observações que o HMM gera podem ser discretas ou contínuas, a depender da definição das distribuições $b_i(\mathbf{x})$. Adicionalmente, mesmo os estados também podem ser contínuos, mas neste trabalho, apenas os HMMs a estados discretos serão considerados.

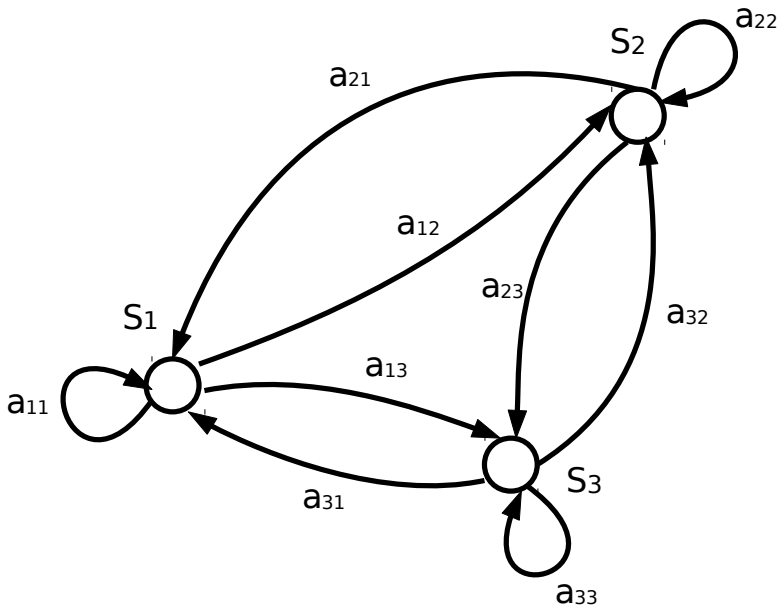


Figura 10 – Uma rede de Markov com três estados $\{S_1, S_2, S_3\}$ e nove probabilidades de transições entre estados $\{a_{11}, a_{12}, ..., a_{33}\}$.

[2] A variável q_t representa que estado está atuando no instante t e o símbolo S_n representa o estado n.

Como classificador, o HMM determina a categoria de uma sequência de observações $\mathbf{X}$ procurando a maior probabilidade $p(\mathbf{X}|\lambda^i)$, sendo o modelo HMM $\lambda = \{A, B, \boldsymbol{\pi}\}$, dentre todas as classes possíveis. $p(\mathbf{X}|\lambda)$ é calculado por meio do algoritmo *Forward–backward* [37]. Considerando a variável de avanço

$$\alpha(i,t) = p(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_t, q_t = S_i|\lambda)$$

isto é, a probabilidade da sequência de observações parcial (até o instante t) e que o estado no instante t seja S_i, dado o modelo λ. Inicializando, a variável de avanço é dada por $\alpha(i,1) = \pi_i b_i(\mathbf{x}_1)$ para $1 \leq i \leq N$, e calculada iterativamente de acordo com

$$\alpha(j,t+1) = \left[\sum_{i=1}^{N} \alpha(i,t)a_{ij}\right] b_j(\mathbf{x}_{t+1}), \quad 1 \leq t \leq T-1$$
$$1 \leq j \leq N$$

o que, por consequência, fornece

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{N} \alpha(i,T).$$

No entanto, o problema mais complicado associado ao HMM é o ajuste dos parâmetro $\lambda = \{A, B, \boldsymbol{\pi}\}$ de forma a maximizar $p(\mathbf{X}|\lambda)$ para um conjunto de dados de treinamento $\mathbf{X}$ [37]. Não há uma forma analítica de encontrar os parâmetros de modo a maximizar a probabilidade da sequência de observações. A forma que será explanada a seguir é chamada de método Baum-Welch (em homenagem a Leonard E. Baum e seus colegas, que estudaram o HMM na década de 1960) [37]. Para tanto, é necessário definir algumas variáveis que auxiliarão no processo de ajuste dos parâmetros. A variável de atraso, complementar à variável de avanço, é definida como

$$\beta(i,t) = p(\mathbf{x}_{t+1}\mathbf{x}_{t+2}...\mathbf{x}_T|q_t = S_i, \lambda).$$

Inicializando $\beta(i,T) = 1$, o processo de indução se dá de forma decrescente em relação a t

$$\beta(i,t) = \sum_{j=1}^{N} a_{ij}b_j(\mathbf{x}_{t+1})\beta(j,t+1).$$

Outra variável necessária é

$$\gamma(i,t) = p(q_t = S_i|\mathbf{X}, \lambda),$$

a probabilidade de estar no estado S_i no instante t, dados a sequência de observações e o modelo. Pode ser calculada a partir das variáveis de avanço e de atraso

$$\gamma(i,t) = \frac{\alpha(i,t)\beta(i,t)}{\sum_{j=1}^{N} \alpha(j,t)\beta(j,t)}.$$

Para o caso do HMM com observações contínuas [3], esse termo torna-se

$$\gamma_t(i,k) = \left[\frac{\alpha(i,t)\beta(i,t)}{\sum_{j=1}^{N}\alpha(j,t)\beta(j,t)}\right]\left[\frac{c_{ik}\mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_{ik},\boldsymbol{\Sigma}_{ik})}{\sum_{m=1}^{M}c_{im}\mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_{im},\boldsymbol{\Sigma}_{im})}\right].$$

Finalmente, a última variável necessária para determinar os parâmetros do HMM pelo método de Baum-Welch é definida como

$$\xi_t(i,j) = p(q_t = S_i, q_{t+1} = S_j|\mathbf{X},\lambda)$$

ou seja, a probabilidade de estar no estado S_i no tempo t e no estado S_j no tempo $t+1$, dados a sequência de observações e o modelo. Novamente, pode ser calculada a partir das variáveis α e β

$$\xi_t(i,j) = \frac{\alpha(i,t)a_{ij}b_j(\mathbf{x}_{t+1})\beta(j,t+1)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha(i,t)a_{ij}b_j(\mathbf{x}_{t+1})\beta(j,t+1)}.$$

O método de Baum-Welch está intimamente ligado ao algoritmo EM [65], usado para estimar os parâmetros da GMM, ou seja, teremos uma sequência de passos para o HMM semelhante ao GMM:

1. Inicializar os parâmetros do modelo $\lambda = \{A, B, \boldsymbol{\pi}\}$;

2. Atualizar os parâmetros seguindo

$$\widehat{\pi}_i = \gamma(i,1)$$

$$\widehat{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma(i,t)}$$

$$\widehat{c}_{im} = \frac{\sum_{t=1}^{T}\gamma_t(i,m)}{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t(i,m)}$$

$$\widehat{\boldsymbol{\mu}}_{im} = \frac{\sum_{t=1}^{T}\gamma_t(i,m)\cdot\mathbf{x}_t}{\sum_{t=1}^{T}\gamma_t(i,m)}$$

$$\widehat{\boldsymbol{\Sigma}}_{im} = \frac{\sum_{t=1}^{T}\gamma_t(i,m)\cdot(\mathbf{x}_t-\widehat{\boldsymbol{\mu}}_{im})(\mathbf{x}_t-\widehat{\boldsymbol{\mu}}_{im})^T}{\sum_{t=1}^{T}\gamma_t(i,m)}$$

[3] Num HMM com observações contínuas, as distribuições de observações por estados são, geralmente, modeladas por uma mistura de gaussianas na forma $b_i(\mathbf{x}) = \sum_{m=1}^{M}c_m\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$.

3. Calcular a log-verossimilhança do modelo atual com os dados de treinamento

$$\log \mathcal{L}(\lambda|\mathbf{X}) = \log p(\mathbf{X}|\lambda) = \sum_{i=1}^{N} \alpha(i,T).$$

Caso não atinga o critério de convergência, voltar ao passo 2.

Novamente, assim como o GMM adaptado com o EM, a estimação dos parâmetros do HMM sofre com uma alta dependência das condições iniciais. A cada iteração, o modelo é aperfeiçoado, isto é, $p(\mathbf{X}|\widehat{\lambda}) \geq p(\mathbf{X}|\lambda)$, onde $\widehat{\lambda} = \{\widehat{A}, \widehat{B}, \widehat{\pi}\}$. No entanto, o algoritmo pode levar a máximos locais no espaço de parâmetros e, por isso, na prática, é executado diversas vezes com condições iniciais diferentes [63].

2.2 Classificação de sons domésticos com GMM

O ponto de partida deste trabalho de pesquisa é o artigo *"Daily sound recognition using a combination of GMM and SVM for home automation"* [4] de Mohamed Sehili e seus colegas [1]. Nele, são comparados dois métodos de reconhecimento de sons, um mais tradicional, baseado num classificador GMM, e outro baseado numa mistura de técnicas, dentre elas a estrutura do UBM [66], os supervetores de médias do GMM [67] e um classificador SVM [63], chamado de SVM GSL. Um mérito destacável desse artigo foi criar uma base de sons, batizada pelo autor de Sons Diários (ou Sons do dia-a-dia) utilizada para testar os métodos propostos. Esta base de dados é composta por 18 classes diferentes (descritas detalhadamente na tabela 1) de sons comuns em ambientes domésticos, amostrados a uma taxa de 16 kHz e armazenados em arquivos wave. Para um ouvinte que não conhece a relação de classes dessa base de sons fica difícil classificar os arquivos apenas com o conhecimento geral, pois, como exemplo, a classe de Respiração pode ser confundida com som de vento ou ruído de fundo, e as classes de Barbeador Elétrico e de Secador de Cabelo podem ser confundidas com sons de motores. No entanto, conhecendo a relação de classes e ao se acostumar com o padrão do sons da base, um ouvinte atento é capaz de classificar arquivos não rotulados com alta precisão.

A primeira estratégia de reconhecimento de sons dessa base de dados, descrita em [1], é bastante direta: a base de dados é dividida em três partes iguais, uma parte é usada para treinar o GMM (com 25-50 componentes por classe) e outra parte é usada para testar o GMM (a parte restante é usada na segunda estratégia, que não é detalhada aqui). De cada arquivo da base são extraídos 16 coeficientes MFCC em janelas de 16 ms do sinal. O *score* de um sinal não rotulado é calculado a partir do GMM θ^i, da i-ésima classe, e da matriz de características $\mathbf{X}$, de ordem $16 \times T$. Ou seja, calcula-se o *score* como

$$p(\mathbf{X}|\theta^i) = \prod_{t=1}^{T} p(\mathbf{x}_t|\theta^i)$$

[4] Reconhecimento de sons diários usando uma combinação de GMM e SVM para automação doméstica.

Tabela 1 – Descrição das classes da base de dados.

Classe	Rótulo	Arquivos	Duração total (s)
Respiração	1	50	106.44
Tosse	2	62	181.69
Louças	3	98	303.77
Porta batendo	4	114	62.70
Porta abrindo	5	21	138.94
Barbeador elétrico	6	62	420.33
Choro feminino	7	36	268.19
Grito feminino	8	70	216.83
Vidro quebrando	9	101	99.52
Secador de cabelo	10	40	224.86
Palmas	11	54	218.65
Balançar de chaves	12	36	166.34
Gargalhadas	13	49	272.65
Grito masculino	14	87	202.11
Papel amassando	15	63	330.66
Espirro	16	32	51.67
Água	17	54	484.72
Bocejo	18	20	95.87

A segunda estratégia, denominada SVM GSL, usa a terceira parte da base de dados para criar um UBM com 1024 componentes. Cada sinal não rotulado gera um supervetor, por meio de uma adaptação *Maximum a Posteriori* (MAP) do UBM, com todos os centros das gaussianas empilhados, isto é, um vetor de dimensão $16 \times 1024 = 16384$. Este supervetor é classificado via SVM, num esquema de um contra um.

A primeira estratégia conseguiu uma taxa de reconhecimento de 69%, enquanto a estratégia SVM GSL conseguiu uma taxa de reconhecimento de 75%. Na presença de ruído, a estratégia do GMM gerou uma taxa de 42% de reconhecimento e a estratégia do SVM GSL 55%, a um SNR de 5 dB. O autor ainda sugere um esquema de classificação hierárquica, utilizando características mais simples num classificador de nível mais baixo — pois os sinais de sons domésticos são bastante distintos — para que em seguida os classificadores mais sofisticados possam resolver os sinais mais semelhantes. No entanto, esse esquema de classificação não foi testado no artigo [1].

Novamente, é interessante notar que as estratégia apresentadas para o reconhecimento de sons domésticos não são ajustadas de uma forma específica para esses tipos de sinais, isto é, a classificação de sons domésticos não aparenta ser essencialmente diferente do reconhecimento de fala/orador.

O estudo do artigo de Sehili et al. [1] faz parte do desenvolvimento do projeto de monitoramento doméstico SWEET-HOME [68], com o objetivo de proporcionar uma tecnologia de interação baseada em áudio que permita ao usuário um controle total sobre

seu ambiente familiar e que identifique situações perigo, facilitando a inclusão social dos idosos. O monitoramento de áudio corresponde a uma parte importante do projeto, que também foca em monitoramento de vídeo, detecção de quedas, sensores fisiológicos (frequência cardíaca, pressão arterial etc) e sensores infravermelhos [69].

3 RESULTADOS EXPERIMENTAIS

3.1 Roteiro de pesquisa

De forma sintética, este trabalho tem como objetivo tentar capturar a informação discriminativa que a dinâmica espectral supostamente carrega, focado em sinais de sons domésticos de uma base de dados criada por Sehili et al. [1]. Além de utilizar a base de sinais de [1], seus métodos e experimentos também são utilizados como patamar inicial para este trabalho, pois eles usam uma abordagem que pouco captura a informação de dinâmica espectral. Isso se deve principalmente ao uso do GMM de forma ingênua. O termo ingênuo se aplica aqui em sentido análogo ao dos classificadores de Bayes (i.e. *Naive Bayes Classifiers*), que são classificadores probabilísticos baseados no teorema de Bayes [70] mas que assumem independência entre observações. No caso do GMM, é assumida independência entre vetores MFCC observados ao longo do tempo.

A representação de um sinal de áudio — via matriz de características, composta de vetores MFCC — ainda deve conter muita informação de dinâmica espectral (apesar de uma parte da informação ser descartada durante a extração do MFCC). Uma forma de imaginar essa dinâmica que se busca capturar é assumindo que cada vetor MFCC é um ponto no espaço, e que a sequência de pontos observados forma uma estrutura geométrica cuja aparência depende da classe de sinal sonoro. Se essa dependência existe, então é porque há informação dinâmica que poderia ser aproveitada na classificação dos sons. Além disso, cada vetor MFCC (calculado a partir de uma janela do sinal original) deve conter em si uma pequena fração de informação de dinâmica espectral, devido ao efeito do princípio da incerteza da Transformada de Fourier [71], que se manifesta através da escolha do tamanho da janela analisada e na taxa de sobreposição. Por exemplo, se uma janela muito larga fosse escolhida para análise – digamos de 1 segundo, com uma taxa de amostragem de 16000 amostras por segundo –, a transformada discreta de Fourier permitiria, portanto, uma resolução espectral de 1 Hz, isto quer dizer que a própria representação espectral do sinal já incorporaria padrões de variação da ordem da percepção humana, como da fala, por exemplo.

De acordo com Giannakopoulos e Pikrakis [61], é possível classificar sinais sonoros entre as categorias de sons Musicais e sons de Fala, utilizando o desvio padrão da sequência temporal do segundo coeficiente MFCC como vetor de características, com um erro Bayesiano estimado de 11,8%. Com o propósito de confirmar essa conclusão, ao reproduzirmos o experimento (Figura 11), usando a base de dados GTZAN[1] [72] de discriminação de

[1] Disponível em <http://marsyasweb.appspot.com/download/data_sets/>.

música/fala, constatamos que o mesmo efeito pode ser realizado tanto pelo segundo coeficiente MFCC quanto pelo quarto e quinto coeficientes, com erros Bayesianos estimados de 14,84%, 8,59% e 11,72%, respectivamente. Esse resultado exibe o alto poder discriminativo dos MFCC, principalmente em relação à variação temporal.

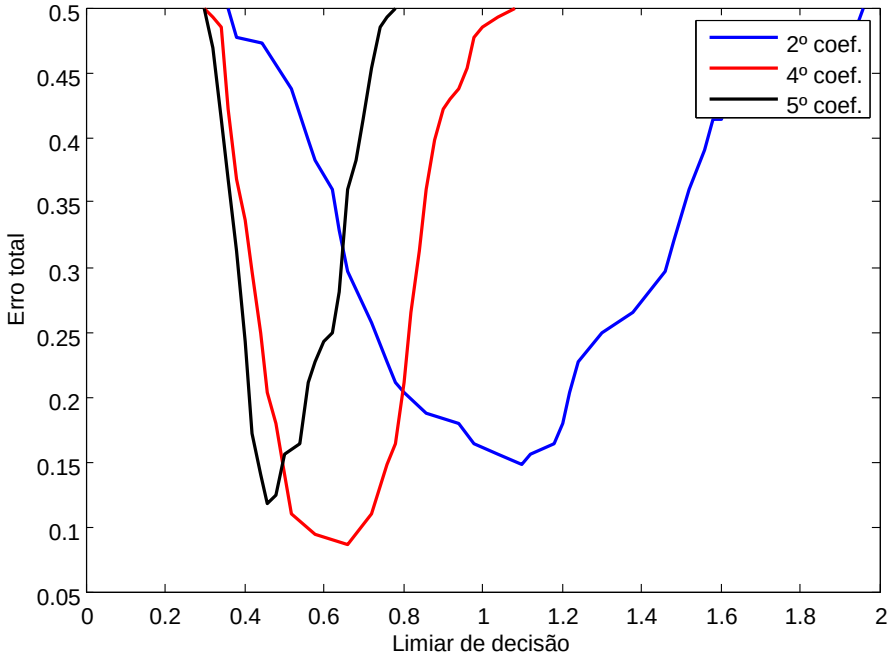


Figura 11 – Erro de classificação em função do limiar de decisão no problema de discriminar Música e Fala, usando como característica apenas o desvio padrão da série temporal de um coeficiente MFCC.

Levando em conta, então, que a maior parte da informação de dinâmica espectral na matriz de características MFCC está na sequência cronologicamente ordenada dos vetores, podemos considerar que as estratégias de classificação dos sinais de sons domésticos de Sehili et al. [1] usam pouca informação dinâmica/rítmica dos coeficientes cepstrais ao longo do tempo, pois pouco importa a ordem dos vetores quando são classificados pelo GMM, segundo o método que reproduzimos de Sehili. Essa limitação é estudada mais em detalhe na próxima seção, onde o trabalho de Sehili et al. [1] é reproduzido passo a passo, como marco zero de incorporação de dinâmica espectral para que, em seguida, sejam analisados métodos de incorporação da informação de dinâmica espectral na mesma base de dados.

O primeiro método de incorporação da dinâmica espectral analisado é o uso dos coeficientes de regressão linear na matriz de características MFCC, conhecidos como delta MFCC [25], gerando um acréscimo na dimensão dos vetores de características, no intuito de anexar informação de evolução temporal dos vetores. Em seguida será analisado o uso do HMM como classificador, visto que é um método tradicionalmente usado em sistemas de reconhecimento de som para capturar a dinâmica dos sinais. Por último, será analisado um esquema de classificação misto, em que um extrator de características que dá ênfase às

modulações de energia por banda, proposto neste trabalho, é combinado ao método de classificação de Sehili et al. [1] no nível de decisão.

3.2 Esquema de classificação básico, reproduzido de Sehili et al. [1]

São dois os pontos críticos para a reprodução dos resultados obtidos por Sehili e seus colegas (cujo esquema de classificação é exposto na figura 12): o método de extração dos MFCCs e o método de classificação baseado no GMM. Apesar de ambos os métodos estarem muito bem descritos na literatura, por diversas fontes [25,33,62], transformar tais métodos em algoritmos computacionais é uma tarefa que exige soluções práticas suplementares. Por causa disso, existem diversos pacotes computacionais disponíveis que implementam métodos e ferramentas comuns em processamento de sinais e reconhecimento de padrões. Mais além, é possível encontrar diversos códigos abertos, em várias linguagens de programação diferentes, disponibilizados por seus autores para uso particular ou em pesquisa. A utilização destes pacotes computacionais fechados é de fácil manejo, principalmente para o usuário inexperiente, no entanto, talvez não seja a melhor alternativa para aquele usuário que deseja ter acesso a aspectos mais específicos dos algoritmos, assim como ao poder de editá-los.

Apesar do trabalho de Sehili et al. [1] ter apresentado duas estratégias distintas de classificação de sons domésticos (a primeira baseada em GMM e a segunda baseada em SVM), apenas a primeira estratégia foi reproduzida e tomada como ponto de partida deste trabalho. A razão disto é que a estratégia baseada em SVM inclui diversas etapas (como a utilização de UBM e supervetores) que dificultam a análise em relação ao efeito da dinâmica espectral na tarefa de classificação.

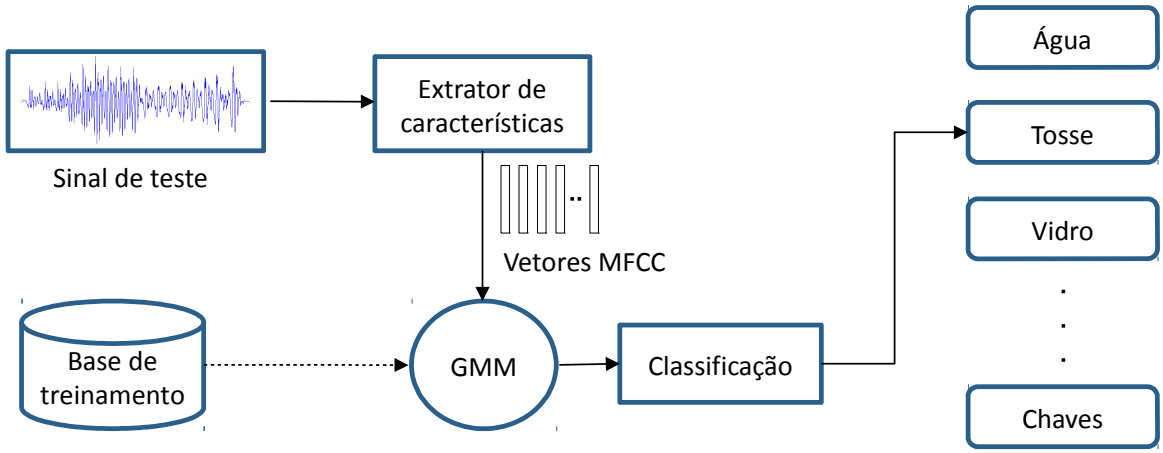


Figura 12 – Primeira estratégia de classificação de Sehili et al. [1], baseada em GMM e MFCC.

No trabalho de Sehili et al. [1], os MFCCs foram calculados a partir de uma rotina computacional desenvolvida pelo próprio autor. Dessa forma, para evitar possíveis discrepâncias nos coeficientes MFCC que poderiam gerar, em reproduções independentes

dos experimentos, resultados não compatíveis com os resultados do artigo, os vetores MFCC usados no trabalho original foram disponibilizados pelo autor para a reprodução dos resultados neste trabalho de mestrado.

Já a rotina computacional do GMM adaptado através do EM foi desenvolvida de forma completamente autônoma, a partir dos textos clássicos e de materiais extras desenvolvidos pelo grupo de pesquisa BioChaves[2] da Universidade Federal de Sergipe, coordenado pelo professor Jugurta Montalvão, do qual o autor deste texto faz parte. Uma etapa decisiva do algoritmo é a inicialização dos parâmetros. É necessário um esquema de inicialização aleatória, nesse problema de classificação, para que se possa analisar a variação da taxa de reconhecimento para diferentes inicializações. Não há consenso no esquema de inicialização do GMM adaptado através do EM [73]. Por exemplo, nesta etapa de reprodução do trabalho de Sehili, foram testados três tipos de inicialização diferentes dos parâmetros do GMM, são elas:

1. **Inicialização aleatória dentro do hiper-retângulo definido pelos dados:** Os vetores de treinamento ocupam uma região do espaço de características que é englobado por um hiper-retângulo, definido pelos valores máximos e mínimos de todos os vetores em cada dimensão. Os M centros da mistura de gaussianas são escolhidos de forma aleatória, a partir de uma distribuição uniforme dentro desse hiper-retângulo. As matrizes de covariâncias são escolhidas como se cada centro gerasse todos os vetores de treinamento, ou seja, cada matriz de covariâncias é calculada como covariâncias amostrais de todos os dados de treinamento da classe, a partir de cada centro. Os coeficientes de peso c_m são inicializados uniformemente, isto é, $c_m = \frac{1}{M}$, em que $1 \leq m \leq M$.

2. **Inicialização por escolha aleatória dos vetores de treinamento:** Cada centro da mistura de gaussianas é escolhido como um dos vetores de treinamento, tomando o cuidado de não escolher vetores iguais para centros diferentes (i.e. sorteio sem reposição). As matrizes de covariâncias e os coeficientes de peso são escolhidos de forma idêntica ao esquema anterior.

3. **Inicialização a partir do *k-means*:** O *k-means* é um método de agrupamento de dados e quantização vetorial no qual o espaço de características é dividido em K células de Voronoi, definidas pelos centros e pela função de distância escolhida [62]. O algoritmo do *k-means* é útil na inicialização do GMM porque ele fornece de imediato regiões de possível aglomeração dos vetores de características pelos centros encontrados.

Foram realizados aqui dez testes na base de dados para cada tipo de inicialização, seguindo a rotina original, isto é, usando um terço da base para treinamento e outro terço

[2] <http://www.biochaves.com/index.htm>

da base para avaliação. Os resultados dos testes estão expostos na figura 13, no qual as taxas de reconhecimento são calculadas considerando o índice total de acertos em relação aos sinais de teste. Os três tipos de inicialização dos parâmetros do GMM não aparentam gerar resultados muito diferentes em relação às taxas de reconhecimento. No entanto, a inicialização a partir dos *k-means* gera modelos de mistura de gaussianas mais bem adaptados aos dados de treinamento que os outros dois tipos de inicialização. Isso pode ser verificado através da avaliação da verossimilhança dos parâmetros do GMM; por causa da boa estimativa inicial dos parâmetros que o *k-means* oferece, os GMMs inicializados com o *k-means* partem de um patamar de verossimilhança maior, se comparados aos outros tipos de inicialização aqui testados, como era de se esperar.

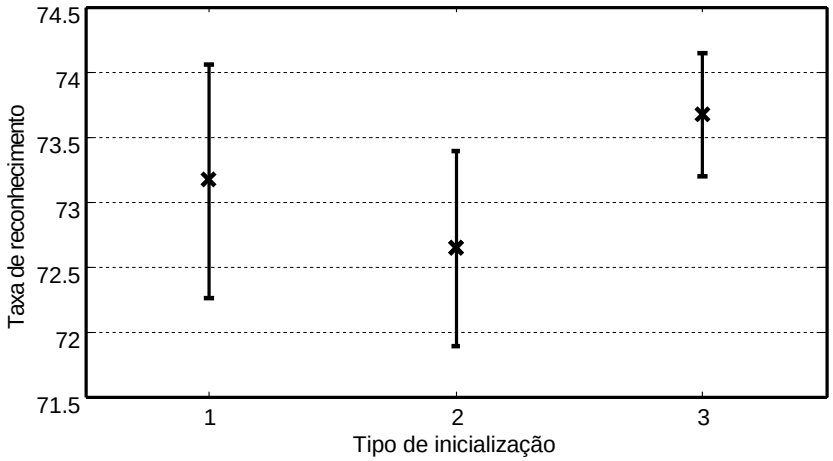


Figura 13 – Taxas de reconhecimento na reprodução do trabalho de Sehili et al. [1], a partir de três tipos de inicializações diferentes do GMM.

Para esses testes foram utilizadas GMMs com 30 componentes ($M = 30$) para cada uma das classes, apesar do trabalho original utilizar um número variável de componentes por classes entre os valores de 25 a 50. Pelo método explicado em [1], todas as classes foram treinadas inicialmente com 50 componentes e, caso o algoritmo do GMM adaptada através do EM não convergisse, esse número seria reduzido gradual e manualmente. As classes com poucos dados de treinamento geralmente estabilizavam com 25 componentes gaussianas. Embora o número de componentes que escolhemos usar seja constante (e não adaptativo, como nos modelos em [1]), as taxas de reconhecimento aqui obtidas são aproximadamente 3% maiores que as taxas obtidas por Sehili et al. [1]. Como os coeficientes MFCCs usados neste experimento foram os mesmo que os coeficientes do artigo original, a discrepância do resultado só pode ser relativa ao treinamento dos modelos de mistura de gaussianas. Apesar da diferença, os resultados foram considerados compatíveis e, portanto, servem como base para os experimentos seguintes a respeito da análise da relevância da dinâmica espectral no reconhecimento de sons domésticos.

A tabela 2 representa uma matriz de confusão obtida com testes nos quais a inicialização dos parâmetros do GMM foi feita pelo *k-means*. O tipo de som que gera mais

confusão na classificação é, sem dúvida, a categoria de Respiração (rótulo 1). Apenas 3 arquivos dessa categoria foram classificados corretamente (de um total de 17), e 10 arquivos foram confundidos com sons de Água (rótulo 17). A categoria de som de Respiração é difícil de ser classificada; no trabalho de Sehili apenas 6 arquivos (de 17) foram classificados corretamente.

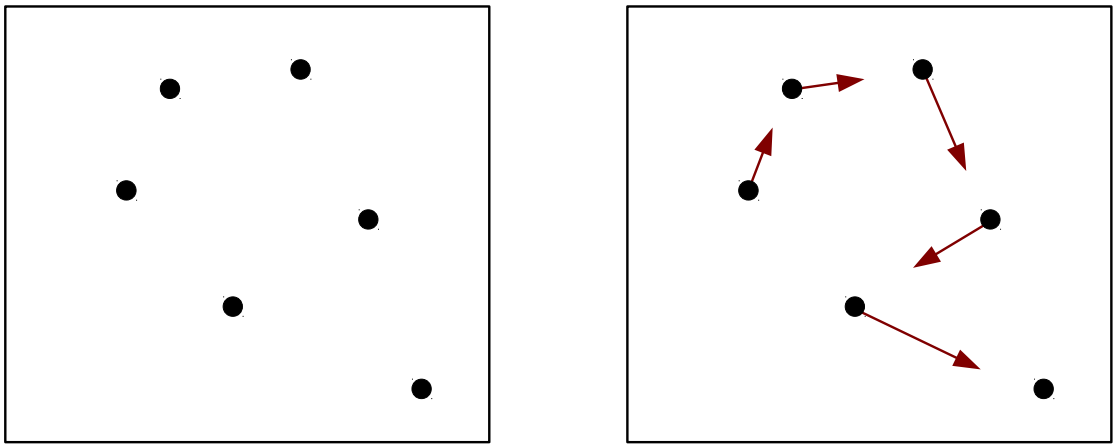
Tabela 2 – Matriz de confusão com taxa de reconhecimento de 74,2%. As classes são rotuladas de acordo com a tabela 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	10	0
2	1	13	0	0	0	0	0	0	0	0	0	0	2	1	4	0	0	0
3	0	0	29	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0
4	1	0	0	36	0	0	0	0	0	0	0	0	0	0	1	0	0	0
5	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	20	0	0	0	0	1	0	0	0	0	0	0	0
7	0	1	0	0	0	0	7	1	0	0	0	0	2	0	0	0	1	0
8	0	0	0	0	0	0	0	18	0	0	0	0	6	0	0	0	0	0
9	0	1	0	0	0	0	0	0	28	0	0	2	0	1	0	0	2	0
10	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0
11	2	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0
13	0	2	0	0	0	0	0	1	0	0	0	0	13	1	0	0	0	0
14	0	0	0	0	0	0	0	5	0	0	0	0	6	17	0	0	1	0
15	0	0	0	0	0	0	0	0	1	0	2	0	0	0	18	0	0	0
16	0	3	0	0	0	0	1	0	0	0	0	0	2	0	1	3	1	0
17	1	0	0	3	0	0	0	0	0	0	0	0	0	0	7	0	7	0
18	0	4	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2

Em 2014, foi realizado um trabalho por Shaukat et al. [74] com o objetivo de investigar de métodos de classificação na tarefa de reconhecimento de sons domésticos. Um dos trabalhos de referência de Shaukat et al. [74] foi o artigo de Sehili et al. [1] e sua base de sinais. Shaukat et al. [74] reproduziu o experimento de Sehili et al. [1] baseado no classificador SVM (chamado de SVM GSL), no qual uma taxa de reconhecimento de 75% foi obtida. Não há nenhum estudo em relação à dinâmica espectral neste trabalho, apesar de vários extratores de características que incorporam informações de dinâmica espectral terem sido usados; o foco do trabalho é a melhora de desempenho no sistema de classificação usando combinações de vários classificadores. Nesse viés, a melhor taxa de classificação obtida por Shaukat et al. [74] (utilizando a mesma configuração experimental de Sehili et al. [1]) é de 77,9%, usando 54 características combinadas e um classificador misto, chamado de *Random Subspace-Random Forest*. Apesar do trabalho de Shaukat et al. [74] utilizar o mesmo artigo de referência deste trabalho de mestrado, não é possível fazer comparativos em relação à reprodução do trabalho de Sehili et al. [1], já que experimentos com focos (e estruturas de classificação) diferentes foram realizados.

3.3 Inserção da informação de variação do MFCC

A maneira possivelmente mais simples de se incorporar informação de dinâmica espectral no vetor de características é incluir a direção de mudança do vetor como característica, algo equivalente ao gradiente numérico, em relação ao tempo, do vetor MFCC. Do ponto de vista do classificador GMM, quando é incorporada a direção de variação do vetor de característica, é como se partes de uma estrutura/trajetória gerada pela dinâmica dos MFCC se tornasse evidente, como na Figura 14b. A Figura 14a ilustra, usando uma hipotética projeção em duas dimensões (para fins de visualização), a forma como os vetores são apresentados ao GMM sem a informação de variação incorporada. Claro que não é possível determinar a evolução temporal dos vetores, ao contrário da Figura 14b em que basta ligar os pontos de acordo com as setas para determinar a trajetória do vetor de características. Quando anexada ao vetor de características, a direção de mudança temporal gera uma ênfase à informação de transição espectral.



(a) Vetores de características de duas dimensões sem a informação de variação temporal.

(b) Vetores de características de duas dimensões com a informação de variação temporal.

Figura 14 – Ilustração do ganho de informação ao incorporar variação temporal aos vetores de características.

Esse tipo de abordagem é chamada na literatura por vários nomes diferentes, entre eles coeficientes de expansão polinomial, coeficientes de regressão linear, derivada temporal, coeficientes delta e, no caso do MFCC, delta MFCC ou ΔMFCC [25, 75]. A maioria dos sistema estado da arte de reconhecimento automático de voz anexam os coeficientes delta e delta-delta (equivalente à segunda derivada) ao vetor de característica para melhorar seu desempenho [76]. Os coeficientes delta podem ser calculados a partir do Quociente de Diferença Simétrica

$$\Delta \mathbf{c}_t = \frac{\mathbf{c}_{t+k} - \mathbf{c}_{t-k}}{2k} \tag{3.1}$$

em que $\mathbf{c}_t$ é um vetor de características de dimensão D no tempo discretizado t, $\Delta \mathbf{c}_t$ o

coeficiente delta e k é o tamanho da janela de análise (na prática $1 \leq k \leq 3$) [25,76,77][3]. A equação 3.1 é o método mais simples de se calcular uma aproximação da derivada numérica, pois não leva em conta os vetores entre $\mathbf{c}_{t-k}$ e $\mathbf{c}_{t+k}$. No entanto, há uma forma de considerar os valores intermediários de k no cálculo do coeficiente delta e generalizar a equação 3.1. Para tanto, considere a matriz $\mathbf{C}_t$ formada pela concatenação dos vetores de característica $\mathbf{c}_{t-k}$ até $\mathbf{c}_{t+k}$, ou seja, $\mathbf{C}_t = [\mathbf{c}_{t-k} \quad \mathbf{c}_{t-k+1} \quad ... \quad \mathbf{c}_t \quad ... \quad \mathbf{c}_{t+k-1} \quad \mathbf{c}_{t+k}]$; cada linha da matriz $\mathbf{C}_t$ representa a evolução temporal em cada dimensão. Analisando-se apenas uma dimensão, pode-se ajustar uma reta $a\mathbf{p}_1 = \mathbf{y} + \boldsymbol{\epsilon}$, de forma que $\mathbf{p}_1 = [-k \quad -k+1 \quad ... \quad k-1 \quad k]^T$, $\mathbf{y}$ seja uma linha da matriz $\mathbf{C}_t$ relativa à dimensão analisada, a seja a inclinação da reta ajustada e $\boldsymbol{\epsilon}$ seja o erro associado ao ajuste. O coeficiente a pode ser estimado por meio do método dos mínimos quadrados como

$$a = (\mathbf{p}_1^T \mathbf{p}_1)^{-1} \mathbf{p}_1^T \mathbf{y}^T. \tag{3.2}$$

Como $\mathbf{p}_1$ é um vetor coluna, então $(\mathbf{p}_1^T \mathbf{p}_1)^{-1} = \left(\sum_{i=-k}^{k} i^2 \right)^{-1}$. Generalizando-se a equação 3.2, de forma que os coeficientes de inclinação de todas as dimensões sejam calculados de uma só vez, e, portanto, a se torne um vetor, agora chamado de coeficiente delta $\Delta\mathbf{c}_t$

$$\Delta\mathbf{c}_t = \frac{\mathbf{p}_1^T \mathbf{C}_t^T}{\sum_{i=-k}^{k} i^2} = \frac{\sum_{i=-k}^{k} i\mathbf{c}_{t+i}}{\sum_{i=-k}^{k} i^2}. \tag{3.3}$$

A equação 3.3 é tão comumente utilizada na literatura para o cálculo de $\Delta\mathbf{c}_t$ quanto a equação 3.1 [50, 78, 79] apesar de cada uma gerar uma estimação diferente da derivada temporal do vetor de características. A Figura 15 ilustra as diferentes estimativas de $\Delta\mathbf{c}_t$ para uma janela de análise $k = 4$, a seta vermelha foi calculada de acordo com a equação 3.3 e a seta azul foi calculada pela equação 3.1.

O coeficiente de segunda derivada temporal $\Delta\Delta\mathbf{c}_t$, seguindo a lógica da equação 3.3, deve levar em conta uma função parabólica ao invés de uma reta, como era o caso de $\mathbf{p}_1$. Assim, o novo vetor deve ter a forma

$$p_{2j} = j^2 - 2(k+1)j + (k+1)^2 - \frac{\sum_{i=1}^{2k+1} i^2 - 2(k+1)i + (k+1)^2}{2k+1}$$

para $j = 1, 2, ..., 2k+1$. Logo,

$$\Delta\Delta\mathbf{c}_t = \frac{\sum_{i=1}^{2k+1} p_{2i}\mathbf{c}_{t-k+1-i}}{\sum_{i=1}^{2k+1} p_{2i}^2}. \tag{3.4}$$

[3] O coeficiente delta-delta $\Delta\Delta\mathbf{c}_t$ pode ser calculado de forma semelhante em cima da sequencia $\Delta\mathbf{c}_t$.

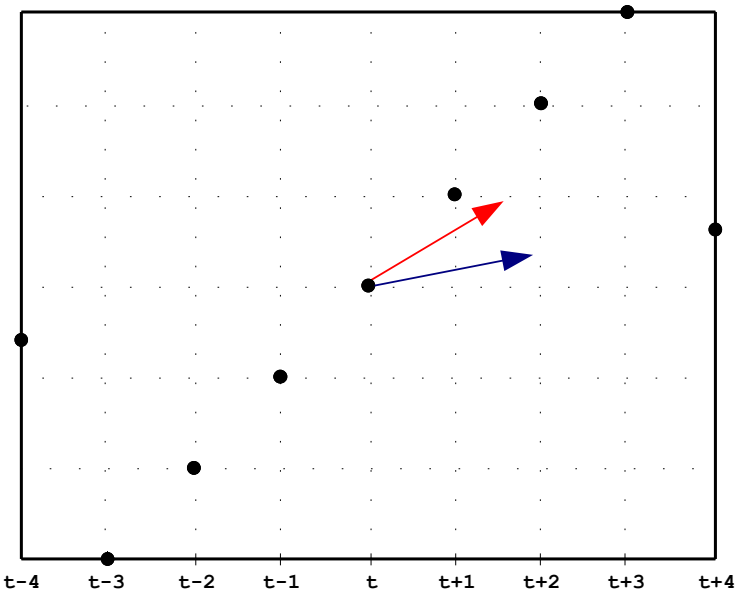


Figura 15 – Ilustração da diferença na estimativa do coeficiente de variação do vetor da posição t em relação às equações 3.1 e 3.3, para $k = 4$.

Finalmente, o vetor de características com os coeficientes delta e delta-delta anexados deve ter a forma

$$\mathbf{x}_t = \begin{pmatrix} \mathbf{c}_t \\ \Delta\mathbf{c}_t \\ \Delta\Delta\mathbf{c}_t \end{pmatrix}. \tag{3.5}$$

Experimentalmente, observa-se que derivadas de ordem mais alta que dois não contêm informação que gere redução no erro de classificação na tarefa de classificação de sons [25]. É importante lembrar que não se deve aumentar a dimensão do vetor de características indefinidamente, devido aos problemas que surgem pelo aumento exponencial do volume do espaço de alta dimensão, conhecido como a Maldição da Dimensionalidade (ou *Curse of Dimensionality*) [62,63]. Apesar disso, um vetor de características formado pelos coeficientes MFCCs, delta e delta-delta deve ter no entorno de 50 dimensões, um valor que não é exageradamente grande.

3.3.1 Testes da influência do ΔMFCC na tarefa de classificação dos sons

São três parâmetros livres, no que se refere à inserção dos coeficientes delta, que podem perturbar o desempenho do sistema de reconhecimento de sons domésticos: o intervalo de análise k, o tipo de cálculo do coeficiente delta (via equação 3.1 ou 3.3) e a utilização do coeficiente de aceleração delta-delta.

Valores comum de k, usados na literatura e em sistemas estado da arte, variam entre 1 e 3 [25, 76, 77], portanto foram esses os valores testados. Os sinais de áudio da base estão amostrados a uma taxa de 16 KHz e os coeficientes MFCCs foram calculados

sobre janelas de 16 ms dos sinais, com sobreposição de 8 ms. Logo, para $k = 1$, a janela de análise temporal dos coeficientes delta será de dois *frames*, portanto de 16 ms. Para $k = 2$, a janela será de 32 ms e, para $k = 3$, a janela será de 48 ms. A figura **??** expõe os resultados dos testes de reconhecimento na base de dados de Sehili et al. [1] para um sistema similar ao descrito na seção 3.2, ou seja, um classificador GMM com 30 centros por classe, treinado a partir de uma inicialização aleatória dos parâmetros e considerando a matriz de características como uma sequência de observações independentes; cada novo sinal a ser classificado recebe um *score* do GMM de cada classe e a classe vencedora é a de maior *score*. No teste, cujos resultados estão expostos na figura 16, estão anexados aos vetores de características MFCC os coeficientes delta calculados pela equação 3.1, para valores de k iguais a 1, 2 e 3. Nota-se, em relação à figura 13, um acréscimo representativo de 1% na taxa de reconhecimento apenas para $k = 1$.

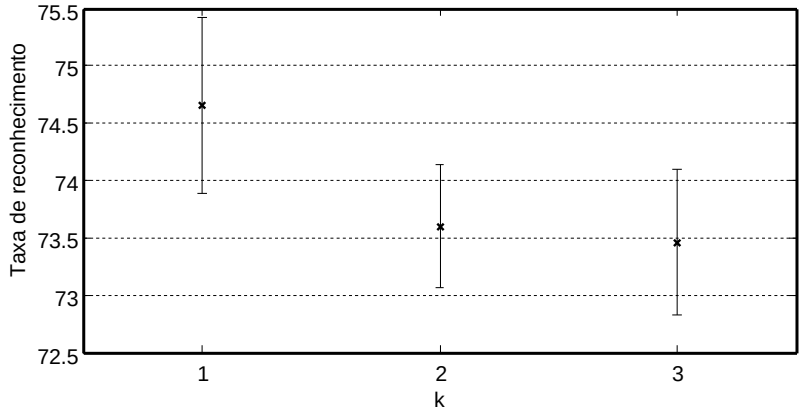


Figura 16 – Resultado dos testes com ΔMFCC a partir da eq. 3.1.

A seguir uma figura semelhante à figura 16, porém com coeficientes delta calculados a partir da equação 3.3. O resultado é semelhante: o único valor de k que parece aumentar razoavelmente a taxa de reconhecimento é $k = 1$. Dessa forma, pode-se concluir que, na formatação deste problema de reconhecimento de sons, os coeficientes delta via equação 3.1 ou 3.3 são equivalentes.

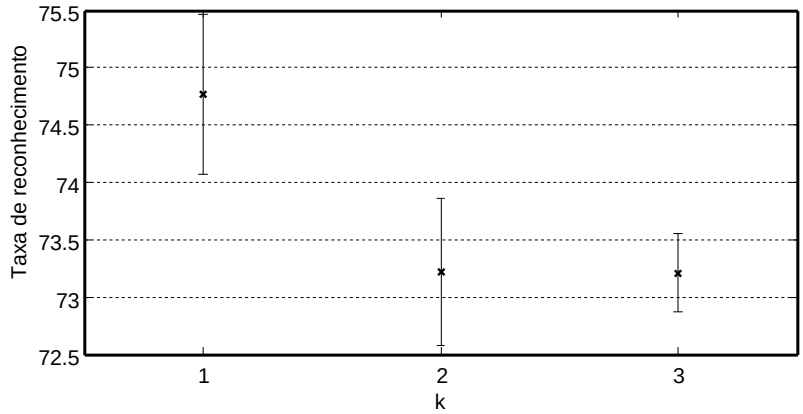


Figura 17 – Resultado dos testes com ΔMFCC a partir da eq. 3.3.

O gráfico da Figura 18, a seguir, faz um comparativo das taxas de reconhecimento em função de k, com os coeficientes delta calculados via equação 3.1 (linha tracejada) ou 3.3 (linha contínua). Esse resultado reforça que a janela de análise de variação dos vetores de características que influencia no desempenho do sistema, e que consequentemente deve carregar informações dinâmicas relevantes, está na ordem de 16 ms. Além disso, pode-se observar novamente a semelhança no resultados que ambos cálculos dos coeficientes delta geram.

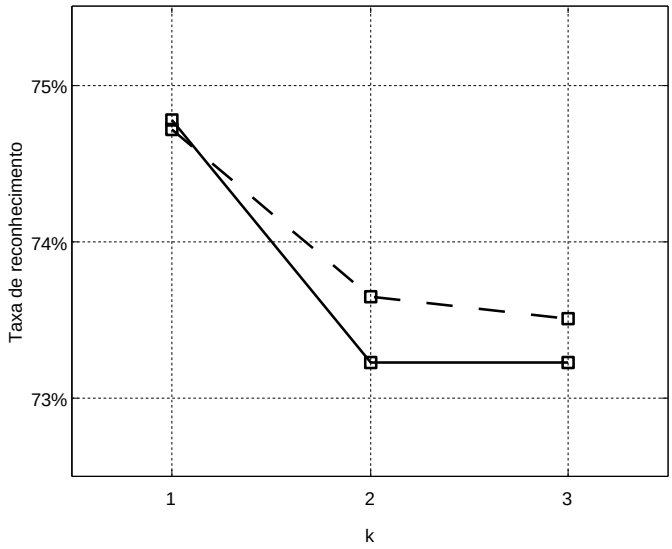


Figura 18 – Desempenho do sistema em função de k com coeficiente delta calculado a partir da eq. 3.1 (linha tracejada) e da eq. 3.3 (linha contínua).

Para o próximo teste, ao vetor de características é anexado tanto o ΔMFCC quanto o $\Delta\Delta$MFCC, como descrito na equação 3.5. Para o cálculo do coeficiente de aceleração, o consenso na literatura é usar $k = 1$ [25, 77, 80], por razões empíricas. Os cálculos dos coeficientes delta-delta foram realizados apenas pela equação 3.1, devido à equivalência, aqui observada, entre os dois métodos de cálculo. Novamente, dez experimentos foram realizados na base de dados, com inicialização aleatória dos parâmetros dos GMMs; os resultados estão expostos na tabela 3. O aumento na média da taxa de reconhecimento desse teste, em relação aos testes com apenas os coeficientes delta anexados, é pouco significativo, no entanto, se comparado aos testes originais de Sehili et al. [1], o aumento é considerável.

Tabela 3 – Resultado dos testes com $\Delta\Delta$MFCC.

Taxa de Reconhecimento (%)										Média
77,1	74,9	76,3	74,3	77,1	73,7	75,7	73,4	75,7	75,1	75,3

Para finalizar, a tabela 4 expõe a matriz de confusão de um teste com os coeficientes delta-delta. Apesar da melhora no sistema, ainda há muito erro de classificação concentrado na classe de Respiração (apenas 3 arquivos foram rotulados corretamente, de 17); e nas classes de Espirro e Bocejo houve mais classificações erradas do que corretas.

Tabela 4 – Matriz de confusão com taxa de reconhecimento de 77,1%. As classes são rotuladas de acordo com a tabela 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	12	0	1	0
2	0	15	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
3	0	0	30	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0
4	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	20	0	0	0	0	1	0	0	0	0	0	0	0
7	0	1	0	0	0	0	7	1	0	0	0	0	3	0	0	0	0	0
8	0	0	0	0	0	0	0	18	0	0	0	0	6	0	0	0	0	0
9	0	1	0	0	0	0	0	0	25	0	0	0	1	0	0	6	1	0
10	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0
14	0	0	0	0	0	0	0	4	0	0	0	0	6	18	0	0	1	0
15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	20	0	0	0
16	0	3	0	0	0	0	1	0	0	0	0	0	4	0	0	2	1	0
17	1	3	0	3	0	0	0	0	0	0	0	0	3	0	2	0	6	0
18	0	3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3

3.4 Sistema com classificador baseado em Modelo Oculto de Markov

O HMM é uma ferramenta poderosa, utilizado em problemas de reconhecimento de padrões (como fala, assinatura, gestos), síntese de fala, linguística computacional, finanças, meteorologia, biologia, criptoanálise, séries temporais, entre muitas outras áreas. Parte da sua popularidade se deve à variabilidade de encaixe do modelo ao problema: são N estados ocultos (na sua versão provavelmente mais usada em aplicações práticas, a versão a estados discretos), cada estado com uma distribuição de probabilidade de observação diferente (seja contínua ou discreta). Dentre alguns dos recursos consequentes do uso do HMM, estão: a possibilidade de gerar sequências de observações a partir de um modelo estocástico dado por um HMM λ, ajuste de um modelo a partir de sequências de observações, avaliação da verossimilhança, $p(\mathbf{X}|\lambda)$, de uma sequência de observações $\mathbf{X}$ dado um modelo λ, inferência dos estados ocultos de uma sequência de observações, comparação direta de modelos HMM. A mesma quantidade de graus de liberdade do HMM que provê tantos recursos, é também causa da explosão do número de parâmetros livres, dificultando a manipulação dos HMMs, e é aí que entram as heurísticas em relação ao sinal modelado [37].

Até então, na procura pelas relações entre a dinâmica espectral dos sinais de som e a classificação de sons domésticos, foi analisado um sistema de classificação considerado de mínima captura de informação de dinâmica espectral, na seção 3.2, e, em seguida,

foi analisado um sistema de classificação que incorpora informações de variação local dos vetores de características, na seção 3.3. A proposta de análise da informação de dinâmica espectral com o HMM é verificar as dependências temporais dos vetores de características através, principalmente, do padrão das probabilidades de transição dos vetores de características entre os centros dos GMMs.

Num sistema dinâmico estocástico, o estado atual depende dos estados anteriores de forma probabilística, ou seja

$$P(q_t = S_i | q_{t-1} = S_j, q_{t-2} = S_k, ...).$$

Quando o conhecimento do estado anterior torna redundante a informação sobre todos os demais estados, este sistema se torna uma Cadeia de Markov. Apesar da Cadeia de Markov poder desprezar as dependências diretas do estado atual com os estado não imediatamente anteriores, a dependência $P(q_t = S_i | q_{t-1} = S_j)$ cria uma corrente que irá ligar toda a sequência de estados pelos quais o sistema dinâmico passar. Essa é a característica principal que o HMM carrega para análise de sinais sequenciais, o padrão de variação do sinal (dinâmica) será determinado pelas probabilidades de transição $P(q_t = S_i | q_{t-1} = S_j)$ do HMM.

De forma propositalmente análoga aos experimentos com GMM, os experimentos de reconhecimento dos sinais da base de sons domésticos com o HMM foram organizados da seguinte forma:

- Cada classe foi representada por um HMM;
- Cada vetor de característica MFCC é uma observação contínua de um HMM;
- Os estados ocultos do HMM foram associados aos rótulos (índices) das gaussianas dos GMMs treinados, como nas seções 3.2 e 3.3. Nos experimentos anteriores foram utilizados misturas com 30 componentes, portanto cada HMM teve 30 estados ocultos, cada um com sua distribuição de probabilidade de observação definida pela componente do GMM da classe correspondente;
- O HMM de cada classe foi treinado de forma a atualizar apenas a matriz de probabilidade de transição do modelo;
- Cada sequência de observações $\mathbf{X}$ (matriz MFCC) é classificada como a classe i de maior probabilidade $p(\mathbf{X}|\lambda^i)$, dentre todos os modelos λ^i, para $i = 1, 2, 3, ..., 18$.

É importante frisar que o intuito dos experimentos desta seção é verificar as dependências temporais dos vetores de características, por isso o uso do HMM é muito limitado em relação ao treinamento – apenas a matriz de probabilidades de transição é atualizada –, caso contrário, qualquer variação na mistura de gaussianas gerada pelo HMM

poderia explicar uma variação na taxa de reconhecimento, não associada à modelagem da dinâmica de encadeamento de observações. Na verdade, utilizando um HMM com uma distribuição de probabilidade de estado inicial $\boldsymbol{\pi}$ igual aos coeficientes c_m do GMM, uma matriz de probabilidades de transição de estados A em que cada linha é igual aos coeficientes c_m e distribuições de probabilidade de observação definidas pelas componentes do GMM, ou seja

$$\begin{aligned} \pi_j &= c_{m=j} \\ a_{ij} &= c_{m=j} \\ b_j(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{m=j}, \boldsymbol{\Sigma}_{m=j}), \end{aligned}$$

e como a probabilidade $p(\mathbf{X}|\lambda)$ depende da variável α, temos que

$$\begin{aligned} \alpha_1(j) &= \pi_j b_j(\mathbf{x}_1) \\ &= c_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \end{aligned}$$

e, portanto, o somatório da coluna $\alpha_1(i)$ é igual à probabilidade do vetor $\mathbf{x}_1$ em relação ao GMM θ. Isto é, $p(\mathbf{x}_1|\theta)$. A fórmula recursiva de α torna-se, então

$$\begin{aligned} \alpha_{t+1}(j) &= \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(\mathbf{x}_t) \\ &= \left[\sum_{i=1}^{N} \alpha_t(i) c_j\right] b_j(\mathbf{x}_t) \\ &= \left[\sum_{i=1}^{N} \alpha_t(i)\right] c_j b_j(\mathbf{x}_t). \end{aligned}$$

Logo, por indução

$$p(\mathbf{X}|\lambda) = \sum_{j=1}^{N} \alpha_T(j) = \prod_{t=1}^{T} \left[\sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\right] = p(\mathbf{X}|\theta),$$

ou seja, nossas escolhas de inicialização de parâmetros fornece um "falso" HMM inicial, pois, de fato, ele é um modelo idêntico ao GMM treinado para a mesma classe de sons e, portanto, não incorpora nenhuma informação de dependência temporal dos vetores de características. Reforçando, o "falso" HMM inicial é aquele em que as linhas da matriz A são todas iguais à distribuição de probabilidade de estado inicial $\boldsymbol{\pi}$. Ademais, essa inicialização dos HMMs permite uma validação da correção dos códigos, pois os resultados esperados, antes dos ajustes dos modelos, devem ser os mesmos já encontrados nos experimentos com os GMMs.

Para verificar a equivalência de cada HMM inicial com o GMM equivalente, figura 19, estão expostos os resultados desse teste em comparação com os resultados obtidos com os GMMs. Apesar dos valores não serem exatamente os mesmos, são estatisticamente equivalentes; as diferenças podem ser relativas às aproximações numéricas dos métodos.

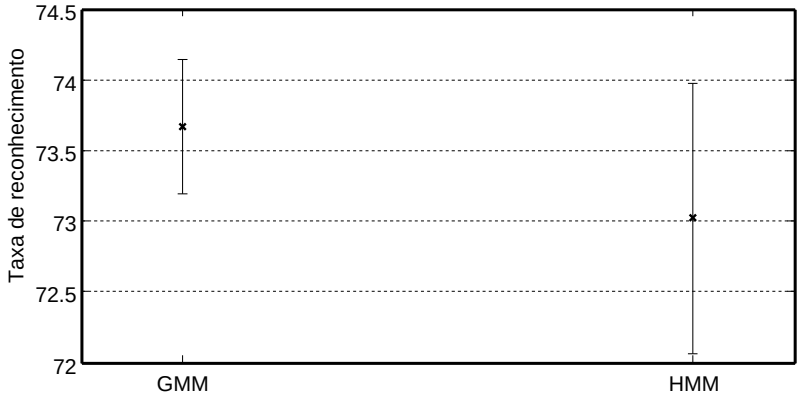


Figura 19 – Taxas de reconhecimento com classificador baseado em HMM sem treinamento em comparação com o classificador baseado em GMM.

Esse mesmo tipo de teste inicial foi realizado a partir dos testes de delta MFCC e, novamente, as taxas de reconhecimento foram compatíveis com os resultados obtidos a partir da classificação por GMM.

Na tabela 5, a seguir, estão os resultados do teste de reconhecimento com a matriz de probabilidades de transição atualizada a partir do método de Baum-Welch e GMMs já treinadas do teste da seção 3.2. As taxas de reconhecimento são consideravelmente mais baixas, um resultado inesperado. Considerar que a dinâmica espectral tem um papel importante no reconhecimento de sons (devido ao seu caráter não estacionário) é uma hipótese heurística naturalmente criada ao se analisar as propriedades e características do som. Essa hipótese é corroborada quando técnicas e métodos, reconhecidos por incorporar e analisar informações espectro-temporais dos sinais, contribuem para o desempenho de um sistema de reconhecimento de sons. No entanto, é possível que o avanço no desempenho de um sistema utilizando tais métodos se deva a razões diferentes da dinâmica espectral, como é o caso do HMM neste experimento. Apesar de, durante a etapa de treinamento do HMM, a verossimilhança dos modelos de cada classe ter aumentado a cada iteração (Figura 20), a taxa de reconhecimento caia. Além disso, se o número de iterações no treinamento aumentar, ocasionando mais aumento da verossimilhança, a taxa de reconhecimento desce mais ainda.

Tabela 5 – Taxas de reconhecimento com classificador baseado em HMM com treinamento da matriz A.

Taxa de Reconhecimento (%)										Média
71,2	70,6	69,8	70,9	71,5	71,5	72	72	71,8	71,8	71,3

O aumento da verossimilhança no treinamento do HMM sinaliza que os vetores de características dos sinais da base têm um padrão de evolução temporal significativo, e que o modelo está capturando esta informação e codificando-a na matriz de probabilidades de transição. Mas, se quanto maior é o ajuste do modelo aos dados, menor é a taxa de reconhecimento, a conclusão à qual se pode chegar é que a informação da evolução temporal dos vetores é, de certa forma, semelhante entre-classes, o que gera confusão na

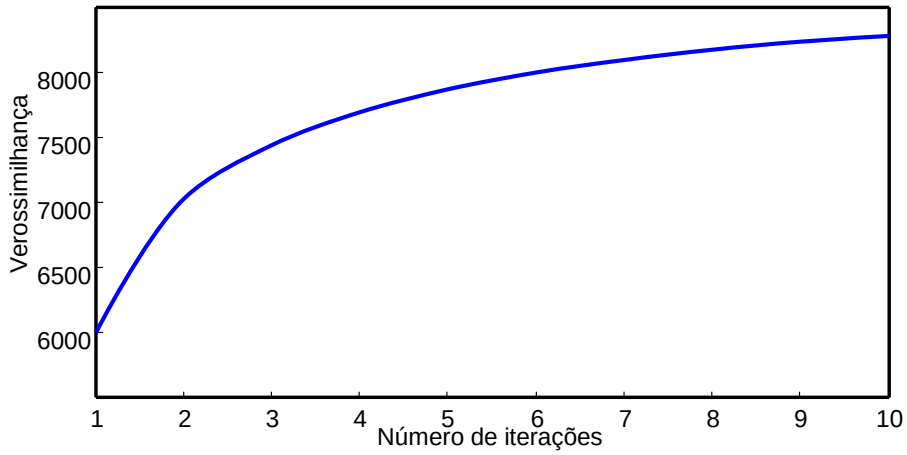


Figura 20 – Variação da verossimilhança média dos modelos das 18 classes em função do número de iterações na etapa de treinamento.

etapa de classificação. De fato, ao analisar a aparência das matrizes A de cada classe, após o treinamento, a maioria tem uma característica em comum: valores altos de probabilidade na diagonal principal, ou seja, tendência de permanência do estado atual. Na Figura 21 estão ilustradas algumas matrizes A após o treinamento do HMM.

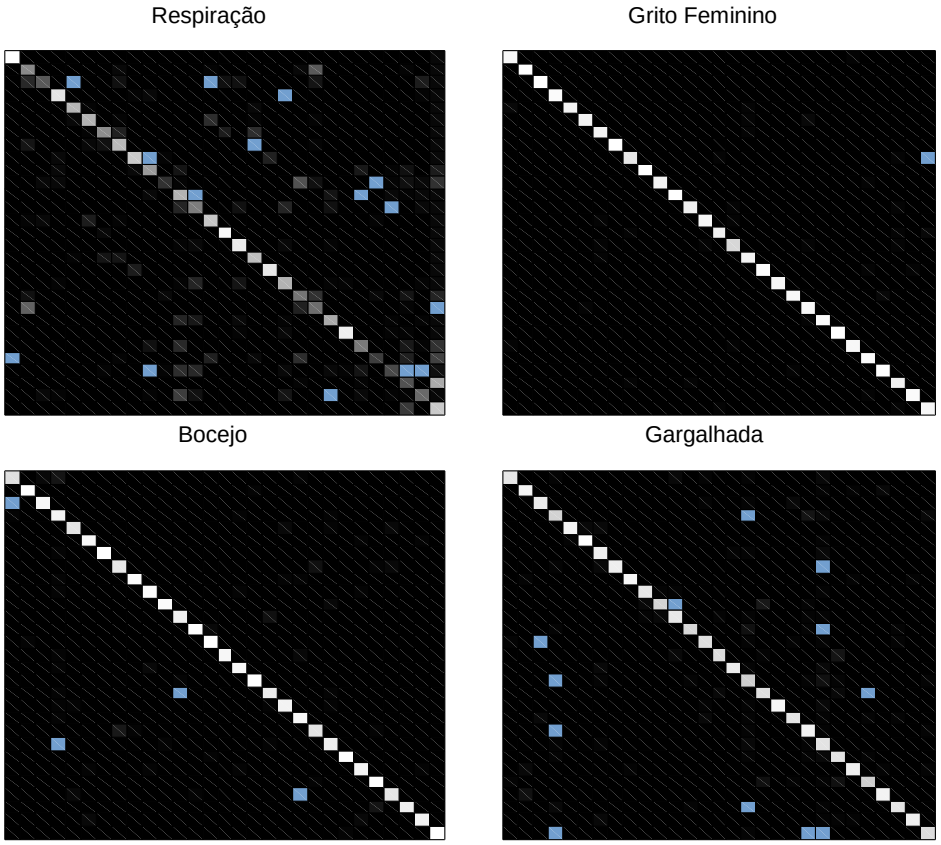


Figura 21 – Algumas matrizes de probabilidades de transição A após o treinamento com o método de Baum-Welch. A característica comum é a concentração de probabilidade na diagonal principal.

O traço da matriz[4] A é um bom parâmetro para verificar a concentração de probabilidade de transição na diagonal principal, ainda mais se for dividido por 30 — valor da soma de todos elementos da matriz A, pois cada linha é uma PMF. A Figura 22 ilustra um gráfico do $\frac{tr(A)}{30}$ para a matriz A de cada classe, com linha contínua. De 18 classes, 11 classes tem $\frac{tr(A)}{30} > 0,6$, de fato a informação que o HMM está aprendendo dos sinais de treinamento não é útil para classificação, na verdade ela só gera confusão. É possível que a janela temporal de análise do HMM ($\approx$ 8 ms, devido ao MFCC) seja pequena e que as transições entre estados diferentes não seja tão constante, ao ponto do número de transições entre estados diferentes ser significativamente menor que as probabilidades de permanência nos estados. Dessa forma, uma solução *ad hoc* para este problema seria subamostrar os vetores de características, no intuito de desprezar boa parte das transições entre estados iguais e destacar transições entre estados diferentes. A título de exemplo, os resultados (linha tracejada da Figura 22) do treinamento do HMM com subamostragem dos vetores de características mostram que a concentração de probabilidade na diagonal principal diminui, gerando matrizes A mais distintas.

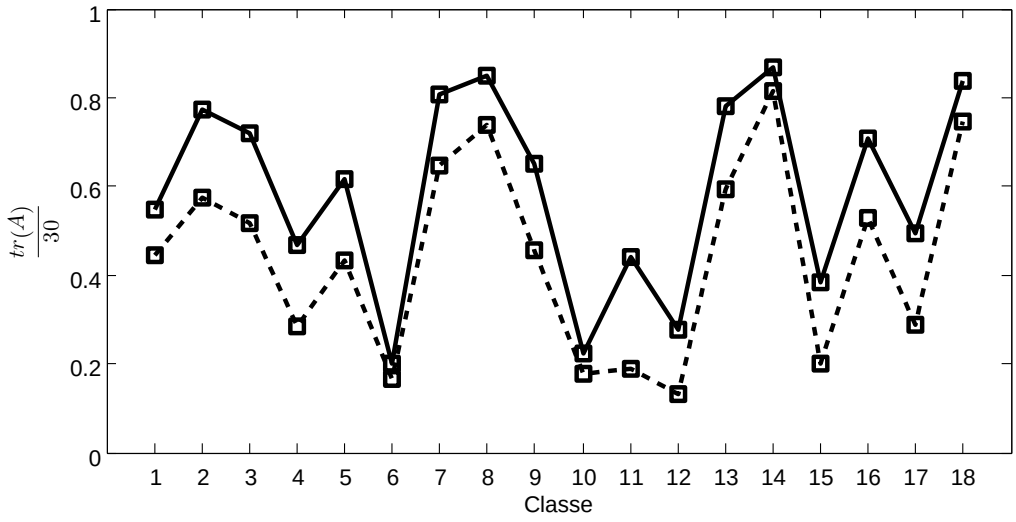


Figura 22 – O traço da matriz A dividido por 30 em função das 18 classes de sons. Linha contínua para matrizes com treinamento tradicional e linha tracejada para matrizes com treinamento a partir de vetores de características subamostrados.

Outros possíveis aprimoramentos para esse sistema de reconhecimento baseado em HMM podem ser adicionados, como a inclusão de uma classe de rejeição quando os valores de $p(\mathbf{X}|\lambda)$ fossem muito baixos, uma análise mais detalhada da variação dos estados pelas observações de cada sinal para melhor mapear a matriz A e até aumentar o número de componentes do GMM por classe no intuito de destacar as transições entre estados. No mais, a conclusão que se chega com esses testes de inclusão do HMM no sistema de classificação de sons domésticos é que a informação de transição entre estados não gera potencial discriminativo entre as classes trabalhadas.

[4] $tr(A) = \sum_i a_{ii}$

Vale ressaltar que o HMM, na forma que foi implementado aqui, está gerando um problema de sobreajuste com os sinais de treinamento. O HMM sem treinamento e atualização da matriz A gera uma taxa de reconhecimento para os dados de teste de aproximadamente 73%±1% (de acordo com a figura 19), enquanto o HMM com treinamento gera uma taxa de reconhecimento de aproximadamente 71,3%±0,7% (de acordo com a tabela 5). Do ponto de vista dos sinais de aprendizado, tanto o HMM sem treinamento quanto o HMM com treinamento classificam corretamente todos os padrões. Mas se o HMM sem treinamento alcança este desemprenho, quem deve estar causando isso é o próprio GMM, já que o HMM sem treinamento foi ajustado para ser um modelo idêntico ao GMM. No entanto, ao observar o comportamento do erro de classificação nos sinais de teste e de aprendizado em função do número de iterações de treinamento para o sistema de classificação baseado em GMM (descrito na seção 3.2), o erro de classificação cai ao longo do treinamento do GMM para os sinais de teste, enquanto que para os sinais de aprendizado, já nas primeiras iterações, o erro de classificação cai para zero. Esse resultado indica que a mudança do modelo GMM para o HMM sem treinamento gerou o sobreajuste, por causa do aumento do número de parâmetros livres principalmente da matriz A. Além disso, há também um indício de que há poucas instâncias de treinamento (pelo menos para algumas classes), o que geraria uma amostragem pobre para a modelagem das classes e, consequentemente, uma estimativa pobre desse modelo, tanto via GMM, quanto via HMM.

Para estudar esse ponto, um pequeno experimento foi realizado para analisar o desempenho do sistema de classificação com um HMM com poucos estados. Num teste com dois estados (tomando o mesmo cuidado de inicializar o HMM de forma semelhante ao GMM da seção 3.2), uma taxa de reconhecimento de 72,8% foi obtida. Num teste com três estados, uma taxa de reconhecimento de 58,7% foi obtida. Tanto no HMM de dois estados quanto no HMM de três estados, o mesmo padrão de concentração de probabilidades na diagonal principal da matriz de transições, observado no HMM de 30 estados, foi verificado. Ou seja, apesar desses experimentos não sofrerem com os problemas advindos do alto número de parâmetros livres do HMM, nota-se que os modelos ajustados também não são capazes de gerar uma boa generalização a partir dos sinais de aprendizado.

3.5 Padrões de Envelope de Energia por Banda

Após realizar diversos testes de reconhecimento de sons domésticos com um sistema baseado naquele proposto por Sehili et al. [1], chega o momento de mudar a abordagem de classificação do sistema, mais precisamente na etapa de extração de características. Até aqui, os MFCCs vêm sendo utilizados como forma de representação dos sons; um extrator de características que analisa *frames* do sinal para capturar um contorno espectral, o que pode ser interpretado como uma análise vertical no plano tempo-frequência. Complementarmente,

um extrator de características que desse ênfase a uma análise horizontal do plano tempo-frequência seria ideal para destacar informações de dinâmica espectral dos sinais.

Sendo assim, nesta seção é proposto um extrator de características para sinais sonoros baseado nos conceitos de Dudley [4] (ver seção 1.2) batizado Padrões de Envelopes de Energia por Banda (PEEB). A proposta desse extrator de características é capturar a informação que os padrões de variação de energia carregam, em várias bandas estreitas do sinal. Para tanto, foi desenvolvido um método de extração de características que segue as seguintes etapas:

1. O sinal $s(t)$ passa por um banco de N_f filtros lineares passa-faixa, com espaçamento igual em escala Mel [81]. Os filtros, para $N_f = 10$, têm suas respostas espectrais ilustradas conjuntamente na Figura 23;

2. Em seguida, é capturado o envelope de energia dos sinais de saída de cada filtro, a partir de uma janela de 229 amostras (escolha empírica), com deslizamento de uma amostra;

3. Cada envelope de energia é subamostrado de 16 kHz para 62 Hz, pois, segundo Dudley [4], a informação essencial para a cognição de sinais de fala nesses envelopes de energia esta encapsulada em padrões de modulação numa escala temporal maior que $\frac{1}{25}$ segundo;

4. Em seguida, os envelopes de energia subamostrados são simbolizados a partir da ordenação crescente de uma janela deslizante de n amostras. Esse processo de simbolização transforma qualquer série temporal numa sequência de símbolos, com base em um dicionário de $n!$ símbolos (como consequência da permutação da ordenação de n amostras) [82];

5. Finalmente, é calculado um histograma de símbolos de cada sequência simbólica de cada canal. Todos os histogramas são, então, concatenados, gerando um vetor de alta dimensão (N_f por $n!$) que representa a dinâmica do sinal de entrada $s(t)$.

Na prática, valores úteis da janela de ordenação variam entre $n = 3, ..., 7$; valores maiores que sete geram um alto custo computacional e valores menores que três não produzem símbolos suficientes para uma análise satisfatória. Na implementação computacional aqui desenvolvida foi utilizado $n = 5$, gerando um dicionário de 120 símbolos. O número de canais escolhido empiricamente foi $N_f = 10$, um valor corroborado pela análise de Dudley [4] em sinais de áudio. Os filtros passa-faixa foram construídos na estrutura de filtros FIR, com largura de banda constantes, iguais a 200 Hz. O envelope de energia foi calculado a partir da energia média da janela de 229 amostras do sinal filtrado. Com essas escolhas de parâmetros, o vetor de características resultante tem 1200 dimensões. A

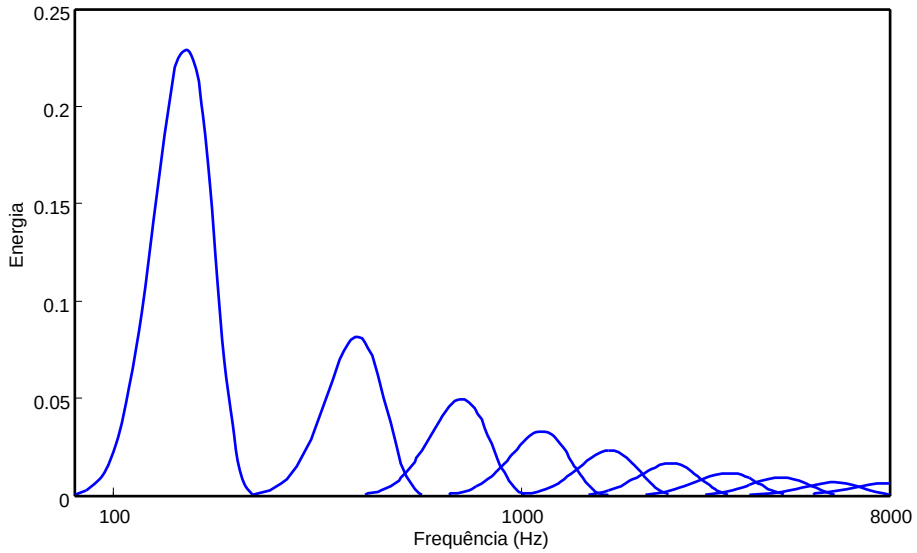


Figura 23 – Dez filtros passa-faixa com os centros espaçados igualmente em escala Mel.

Figura 24 ilustra um modelo esquemático do cálculo desse vetor de característica baseado na simbolização do perfil de energia por banda do sinal.

O extrator de características aqui proposto é, sem dúvida, um protótipo ainda em experimentação, com vários pontos de ajustes ainda por serem estudados mais finamente. Um aspecto importante desse extrator é que, independente do tamanho do sinal de entrada, o vetor de características continuará do mesmo tamanho, diferente da extração típica de MFCCs, por exemplo, em que se calcula um vetor de características a cada *frame* do sinal. Esse é um aspecto que torna muito cômodo os testes de reconhecimento com a base de sons domésticos, já que não é necessário assumir independência entre observações (caso do GMM).

O esquema de classificação usando os PEEB foi organizado da seguinte forma: cada classe é definida por um padrão de referência, calculado como a média de todos os padrões dos arquivos de treinamento da respectiva classe. Cada novo padrão desconhecido é classificado a partir das distâncias euclidianas entre o padrão desconhecido e os padrões de referência de cada classe, ou seja, o padrão de referência mais próximo do padrão desconhecido define a sua classe. Como não há nenhum aspecto aleatório nesse esquema de classificação, isto é, os resultados serão sempre os mesmo para os mesmos arquivos de treinamento e de teste, não há necessidade de repetições no teste de reconhecimento com a base de sons domésticos de Sehili et al. [1].

Sendo assim, o teste resultou numa taxa de reconhecimento de apenas 55,9%. É possível constatar que a informação que os PEEB estão destacando é diferente, e de certa forma complementar, à informação capturada com o esquema de classificação baseado em MFCC, pela verificação da matriz de confusão gerada (tabela 6). A classe Respiração, a mais difícil de ser reconhecida dentre os testes anteriores, teve um resultado um tanto

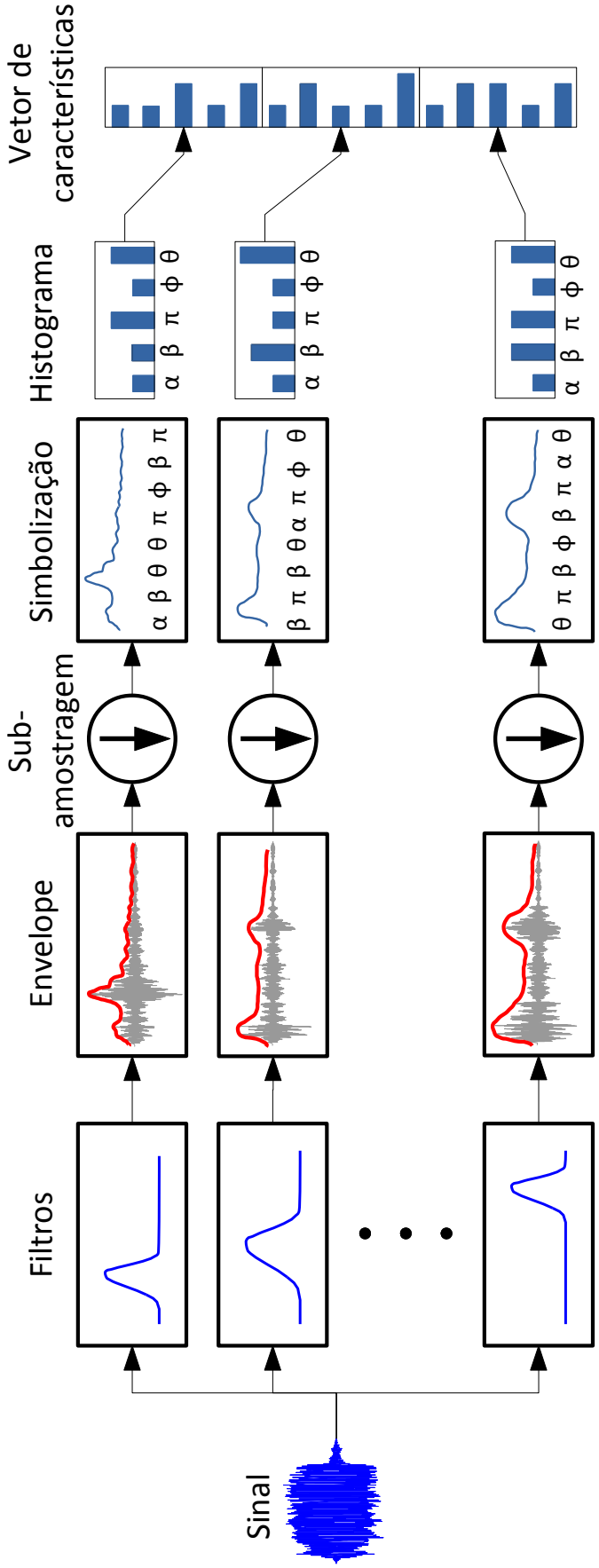


Figura 24 – Esquema do extrator de características baseado em simbolização do perfil de energia por banda do sinal.

Tabela 6 – Matriz de confusão do esquema de classificação baseado em PEEB. Taxa de reconhecimento de 55,9%.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	13	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	1
2	0	9	2	2	0	0	2	0	0	0	1	0	2	0	1	2	0	0
3	0	5	9	5	1	0	0	0	3	0	7	0	0	0	1	2	0	0
4	0	0	2	28	0	1	0	0	4	1	1	0	0	0	0	0	1	0
5	0	0	1	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	16	0	0	0	4	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	11	0	0	0	0	0	1	0	0	0	0	0
8	4	0	0	0	0	2	0	10	0	0	0	1	0	2	2	0	0	3
9	0	1	4	2	0	1	0	0	18	3	0	0	0	0	0	0	4	1
10	0	0	0	0	0	3	0	0	0	11	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0
12	0	1	0	0	0	1	0	0	0	0	0	9	0	0	1	0	0	0
13	0	0	0	0	0	0	11	1	1	0	0	0	3	0	1	0	0	0
14	2	1	0	1	0	9	0	3	0	0	0	0	2	9	2	0	0	0
15	0	1	0	0	1	2	0	1	0	1	0	0	0	0	14	0	1	0
16	1	0	0	2	0	0	2	1	0	0	0	0	1	0	0	3	0	1
17	0	0	0	0	0	1	0	0	0	0	0	5	0	0	2	0	10	0
18	0	0	0	0	0	0	0	1	0	0	0	0	3	2	0	0	0	1

impressionante com os PEEB: 13 arquivos corretos de um total de 17.

Se o MFCC e os PEEB capturam informações distintas dos sinais de som, uma alternativa para melhorar o desempenho do sistema seria combinar essas duas técnicas. Devido às suas diferenças de formato, a opção mais fácil de implementar é combinar os métodos no nível da decisão. Para cada sinal a ser classificado são calculadas 18 *scores* no formato $\log p(\mathbf{X}|\theta_i)$ pela estratégia baseada em GMM e MFCC; já para a estratégia baseada em PEEB são calculados 18 *scores* no formato de distância euclidiana. Para combinar os *scores* a fim de tomar uma decisão de classificação é necessário fazer uma transformação na métrica dos *scores*. A maneira, possivelmente mais simples, seria normalizar ambos *scores* entre os valores de 0 a 1 e inverter um dos *scores*, já que o baseado em PEEB expõe um valor de dissimilaridade e o baseado em MFCC expõe um valor de similaridade. Finalmente, os *scores* são combinados por uma soma ponderada $\rho \cdot scores_{MFCC} + (1 - \rho) \cdot scores_{PEEB}$ e o sinal desconhecido é classificado como o de maior valor dentre as 18 classes. A Figura 25 ilustra um gráfico da taxa de reconhecimento em função do coeficiente de ponderação ρ, aparentemente o valor ideal da combinação é $\rho \approx 0,75$.

Os resultados expostos na tabela 7 são da combinação dos scores PEEB com os GMMs treinados do teste da seção 3.2 para $\rho = 0,75$. Esses resultados mostram os benefícios que o PEEB pode oferecer ao sistema de reconhecimento de sons domésticos com sua ênfase no aspecto de evolução espectro-temporal dos sinais. Foi realizado também um teste combinando o PEEB com o ΔMFCC, de forma análoga ao experimento anterior,

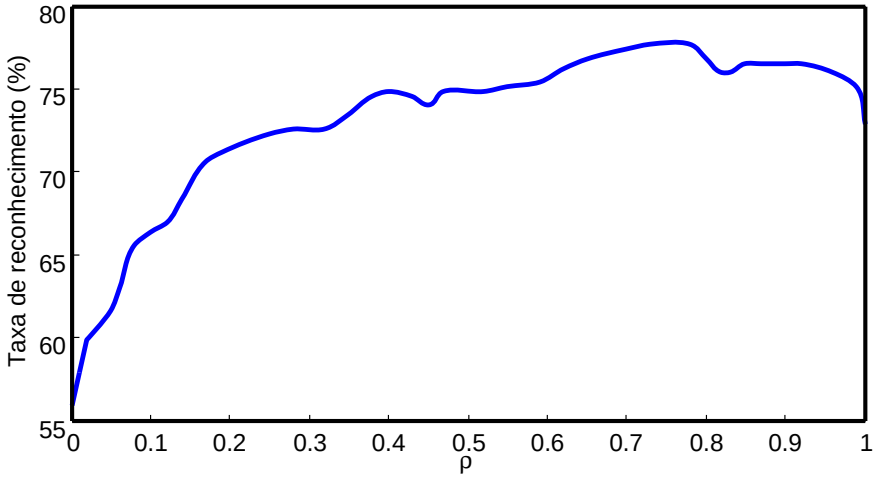


Figura 25 – Taxa de reconhecimento em função do coeficiente de ponderação ρ da combinação de *scores*.

permitindo verificar se essas duas abordagens capturam informações diferentes dos sinais analisados. O resultado desse teste gerou taxas de reconhecimento semelhantes às taxas da tabela 6, isso indica que o PEEB captura informações que incluem as capturadas pelo ΔMFCC.

Tabela 7 – Taxas de reconhecimento com classificador baseado na combinação das técnicas do MFCC, GMM e PEEB, com $\rho = 0,75$.

Taxa de Reconhecimento (%)										Média
78,8	78,0	77,4	77,1	77,7	77,1	77,1	76,8	76,8	77,1	77,4

No trabalho de Shaukat et al. [74], o melhor aperfeiçoamento em relação ao trabalho de Sehili et al. [1] obteve uma taxa de reconhecimento de 77,9%. Como o aperfeiçoamento de Shaukat et al. [74] se deu, principalmente, na etapa de classificação, é possível que usando um esquema de classificação mais bem elaborado com os PEEB o desempenho do sistema aqui apresentado, que combina características PEEB e MFCC, melhore ainda mais.

Para finalizar, um comparativos dos experimentos efetuados nesse capítulo está exposto na Figura 26.

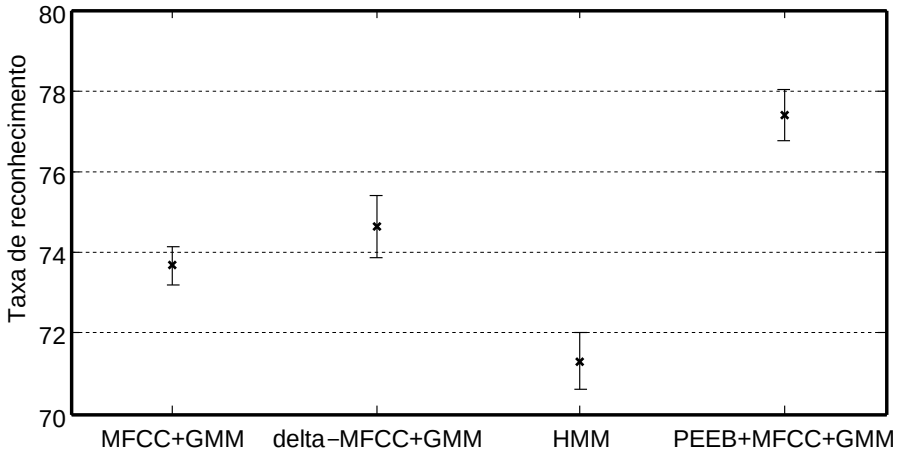


Figura 26 – Comparativos dos sistemas de reconhecimento de sons domésticos testados ao longo deste capítulo.

4 CONCLUSÕES

Neste trabalho, foi estudada a hipótese de que a dinâmica espectral possui regularidades que podem ser modeladas em sons tipicamente domésticos, afetando, por consequência, a tarefa de classificação sonora. Tanto os experimentos de base (de referência) quanto a base de sons domésticos foram reproduzidos e obtida, respectivamente, do trabalho de Sehili et al. [1]. De fato, [1] foi escolhido como ponto de partida, pois representa uma estratégia de classificação cuja ordem dos vetores de características é insignificante, portanto uma estratégia que descarta boa parte da informação de dinâmica espectral do sinal. A proposta de análise da dinâmica espectral neste trabalho foi idealizada de forma que, a cada novo experimento, uma técnica, que dá destaque a um aspecto da dinâmica espectral, fosse incorporada. Além disso, foi tomado o cuidado para que cada nova técnica analisada não incorporasse aspectos alheios ao da dinâmica espectral, o que geraria conclusões equivocadas.

A incorporação dos coeficiente de regressão linear da matriz MFCC foi a primeira técnica testada. Por ser uma técnica tradicional dos sistemas de reconhecimento de sons [25, 73, 75], era esperado que fossem obtidos bons resultados com seu uso, pois é o que se observa frequentemente em reconhecimento de fala. Mais que isso, ela representa uma informação local da dinâmica espectral, pois cada vetor MFCC representa um ponto num espaço de características espectrais, mas do ponto de vista do GMM, uma sequência de vetores MFCC não é nada mais do que uma aglomeração de pontos no espaço. No entanto, cada vetor MFCC cria uma gama de possibilidades para o vetor subsequente baseado em algo equivalente ao seu momento linear. Se, então, a informação de variação temporal dos vetores MFCC for incorporada ao vetor de características, vetores MFCC com posições equivalentes no espaço, mas com derivadas diferentes, serão percebidos diferentemente pelo GMM, possibilitando um maior grau de discriminação entre classes. O uso dos coeficientes de regressão linear gerou um aumento sensível na taxa de reconhecimento de sons domésticos, para o sistema baseado em GMM. Podemos concluir, dessa forma, que a dinâmica dos vetores MFCC carrega uma informação extra (além da sua posição no espaço) para caracterizar cada classes. No entanto, a incorporação da informação embutida na sequência ordenada dos vetores pelos coeficientes delta é apenas local (pois a janela de cálculo é de apenas 16 ms), isso sugere que uma incorporação global da sequência de vetores pode gerar melhores resultados.

Uma técnica que permite a analise de sequências ordenadas de observações é o HMM, também usada tradicionalmente em sistemas de reconhecimento de sons [37, 83, 84] e até no reconhecimento de sons domésticos [85]. A análise estocástica do HMM se concentra na

determinação da matriz de probabilidades de transição entre estados $P(q_t = S_j|q_{t-1} = S_i)$: a sequência ordenada dos vetores MFCC de um sinal de teste deve ter um padrão de variação de estados consequentes à matriz de probabilidades de transição para ser classificado corretamente. Apesar da expectativa em relação ao HMM como modelo mais refinado e classificador mais complexo, os experimentos com o HMM só geraram mais erros de classificação, se comparados aos experimento com ΔMFCC e até aos experimentos de [1]. Foi constatado que a razão da maior taxa de confusão com o HMM se devia à semelhança das matrizes de probabilidades de transição dos modelos entre classes diferentes: a característica principal é a alta probabilidade de permanência num estado. A razão desse problema não foi determinada precisamente, mas duas hipóteses são consideradas. A primeira hipótese é que a taxa de amostragem dos vetores MFCC esteja num valor alto, em relação à taxa de variação de estados diferentes, o que geraria um destaque na repetição de observações de um mesmo estado. Outra hipótese é que o número de componentes da mistura de gaussianas esteja pequeno para uso com HMM – apesar de o GMM estar bem ajustados aos dados, de acordo com os experimentos anteriores –, sabendo da equivalência aqui estabelecida entre o número de componentes do GMM e o número de estados do HMM, mais componentes da mistura (e portanto, mais estados) geraria mais transições entre estados diferentes e, assim, a possibilidade de matrizes de probabilidades de transição mais diversas entre classes. A verificação dessas hipóteses não pôde ser incluída neste trabalho e, portanto, será uma tarefa para trabalhos futuros.

Como conclusão dos experimentos aqui realizados, o HMM provavelmente não foi dimensionado de forma a prover todo o seu potencial na análise da dinâmica dos vetores MFCC e, consequentemente, não desempenhou papel significante na classificação dos sons domésticos. Mas a estrutura escolhida para o HMM, neste trabalho, permitiu concluir que o GMM equivalente ao usado em [1] não se beneficia da inclusão, no modelo, de dependências do tipo Markoviana entre observações.

O último experimento de classificação dos sons domésticos foi baseado no extrator de características chamado de Padrões de Envelope de Energia por Banda (PEEB). Os PEEB representam uma mudança de paradigma em relação aos experimentos anteriores, onde a análise da dinâmica espectral na classificação dos sons com a matriz MFCC depende de uma estratégia de captura da informação escondida na sequência ordenada dos vetores MFCC. Em contraste, com os PEEB os próprios vetores de características já incorporam a informação de evolução espectro-temporal dos sinais. Apesar dos PEEB ainda estarem em fase de evolução, os experimentos de classificação de sons domésticos com combinação das estratégias (MFCC e PEEB) mostraram a possibilidade de ganho em desempenho no sistema. É importante notar que a principal análise de dinâmica com o PEEB está na etapa de simbolização, em que 5 amostras do envelope de energia subamostrado de uma banda de frequência são simbolizados a partir da ordenação dos valores. Sendo assim, como o envelope de energia foi subamostrado de 16 kHz para 62 Hz, e uma sequência de 5

amostras é simbolizada, então a janela de análise temporal é de apenas $(\frac{1}{62}) \cdot 5 \approx 80$ ms, um valor na mesma ordem de grandeza da janela de análise dos coeficientes delta (16 ms). De acordo com os testes da seção 3.5, o PEEB engloba as informações capturadas pelo ΔMFCC.

Com os resultados desses experimentos em mãos, pode-se concluir que a dinâmica espectral desempenha um papel relevante para o reconhecimento de sons domésticos. O horizonte de perspectivas para o futuro é animador, dado o crescente número de pesquisas na área de aprendizado de máquina e reconhecimento de padrões, e também na área de monitoramento de áudio e televigilância sonora [1, 3, 12].

4.1 Trabalhos futuros

Ainda no tópico do estudo da relevância da dinâmica espectral no reconhecimento de sons domésticos, vários experimentos complementares merecem atenção. Um deles é a análise de outros extratores de características que incorporem informações espectro-temporais do sinal, como os coeficientes DWTC [3], o *Relative Spectral (RASTA) Perceptual Linear Prediction* [5], entre outros. Desenvolver o extrator de características PEEB é outra tarefa importante, já que foi mostrado neste trabalho o potencial discriminativo do PEEB. Por exemplo, como proposta, a análise simbólica multicamada pode ser uma alternativa para aumentar a janela temporal de análise do PEEB. Adicionalmente, nota-se que a taxa de classificação é dependente da classe sonora, sugerindo formas mais elaboradas de fusão das características PEEB e MFCC para cada classe.

Como os resultados aqui obtidos com o HMM não foram satisfatórios, dado o potencial do HMM, uma análise mais minuciosa dos problemas de classificação se faz necessária no futuro. Por exemplo, uma proposta discutida na seção 3.4 é o treinamento dos modelos sequências de observações MFCC subamostradas, de forma a dar destaque às transições entre estados diferentes e, consequentemente, gerar matrizes de transições mais distintas entre classes.

Um aspecto importante no desenvolvimento deste trabalho é a expansão da base de sons domésticos criada por Sehili et al. [1], que permitirá a realização de testes e experimentos mais expressivos e condizentes com a realidade. Com o aumento da base, muitos novos sons poderiam ser incluídos, inclusive alguns sem classificações conhecidas. Assim, como ainda não há uma classe de rejeição, seria interessante a inclusão dessa classe, já que muitos sons em um ambiente real também podem não possuir classificações conhecidas pelo sistema. Paralelamente, o desenvolvimento de um protótipo *online* do sistema de classificação de sons domésticos deverá levantar outras questões relevantes quanto aos métodos utilizados e à relevância da dinâmica espectral.

REFERÊNCIAS

1 SEHILI, M. et al. Daily sound recognition using a combination of gmm and svm for home automation. In: IEEE. *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. [S.l.], 2012. p. 1673–1677. 4, 7, 10, 11, 12, 18, 22, 29, 33, 34, 36, 37, 38, 40, 41, 45, 46, 53, 55, 58, 60, 61, 62

2 US, N. C. for H. S. et al. Health, united states, 2008: with special feature on the health of young adults. National Center for Health Statistics (US), 2009. 6, 11, 12

3 CROCCO, M. et al. Audio surveillance: a systematic review. *arXiv preprint arXiv:1409.7787*, 2014. 11, 12, 13, 14, 22, 24, 29, 62

4 DUDLEY, H. Remaking speech. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 11, n. 2, p. 169–177, 1939. 11, 18, 19, 22, 24, 54

5 HERMANSKY, H. Speech recognition from spectral dynamics. *Sadhana*, Springer, v. 36, n. 5, p. 729–744, 2011. 11, 15, 18, 22, 62

6 GIACOMIN, K. C. Dez anos do conselho nacional dos direitos do idoso. *Revista Portal de Divulgação*, n. 26, 2012. 12

7 FUND, N. P.; INTERNATIONAL, U. H. A. *Ageing in the twenty-first century: a celebration and a challenge*. [S.l.]: United Nations Population FundHelp Age New YorkLondon, 2012. 12

8 DOMINGUES, R. G.; FILHO, A. Carlos de P. A domÓtica como tendÊncia na habitaÇÃo. III Simpósio de Pós-Graduação em Engenharia Urbana, 2012. 12

9 OHTA, S. et al. A health monitoring system for elderly people living alone. *Journal of telemedicine and telecare*, SAGE Publications, v. 8, n. 3, p. 151–156, 2002. 12

10 BARLOW, J. et al. A systematic review of the benefits of home telecare for frail elderly people and those with long-term conditions. *Journal of Telemedicine and Telecare*, SAGE Publications, v. 13, n. 4, p. 172–179, 2007. 12

11 VACHER, M. et al. Complete sound and speech recognition system for health smart homes: application to the recognition of activities of daily living. *New Developments in Biomedical Engineering*, In-Tech, p. pp–645, 2010. 12

12 FOGGIA, P. et al. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, Elsevier, v. 65, p. 22–28, 2015. 12, 13, 22, 26, 62

13 MONTALVAO, J.; MONTALVAO, M. V.; RAULINO, C. Deteccao de orador e palavras em televigil^ancia médica com treinamento mínimo: Uma amostra por palavra. 12

14 COMISSION, E. *Home Sweet Home Project*. 2015. Acessado em 23 de setembro de 2015. Disponível em: <http://www.homesweethome-project.be/>. 12

15 SUN, B. P. L. L. J.; VELASTIN, S. A. Fusing visual and audio information in a distributed intelligent surveillance system for public transport systems “. 2003. 13

16 VALENZISE, G. et al. Scream and gunshot detection and localization for audio-surveillance systems. In: IEEE. *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. [S.l.], 2007. p. 21–26. 13

17 PETERSEN, J. K. *Understanding surveillance technologies: spy devices, their origins & applications*. [S.l.]: CRC Press, 2002. 13

18 RABAOUI, A. et al. Using one-class svms and wavelets for audio surveillance. *Information Forensics and Security, IEEE Transactions on*, IEEE, v. 3, n. 4, p. 763–775, 2008. 13, 22

19 COWLING, M.; SITTE, R. Comparison of techniques for environmental sound recognition. *Pattern recognition letters*, Elsevier, v. 24, n. 15, p. 2895–2907, 2003. 14, 29

20 JUNEJA, M.; SANDHU, P. S. Performance evaluation of edge detection techniques for images in spatial domain. *methodology*, v. 1, n. 5, p. 614–621, 2009. 15

21 SCHWARTZ, E. L. et al. Shape recognition and inferior temporal neurons. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 80, n. 18, p. 5776–5778, 1983. 15

22 LEE, H. et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: ACM. *Proceedings of the 26th Annual International Conference on Machine Learning*. [S.l.], 2009. p. 609–616. 15

23 ZOU, W. et al. Deep learning of invariant features via simulated fixations in video. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 3212–3220. 15

24 POTAMIANOS, A.; MARAGOS, P. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 99, n. 6, p. 3795–3806, 1996. 16

25 HUANG, X. et al. *Spoken language processing: A guide to theory, algorithm, and system development*. [S.l.]: Prentice Hall PTR, 2001. 16, 37, 38, 42, 43, 44, 46, 60

26 PARK, C.-y. *Consonant landmark detection for speech recognition*. Tese (Doutorado) — Massachusetts Institute of Technology, 2008. 17

27 ALI, A. M. A.; SPIEGEL, J. Van der; MUELLER, P. Acoustic-phonetic features for the automatic classification of stop consonants. *Speech and Audio Processing, IEEE Transactions on*, IEEE, v. 9, n. 8, p. 833–841, 2001. 17

28 MCGURK, H.; MACDONALD, J. Hearing lips and seeing voices. *Nature*, v. 264, p. 746–748, 1976. 18

29 BISWAS, A.; SAHU, P.; CHANDRA, M. Admissible wavelet packet features based on human inner ear frequency response for hindi consonant recognition. *Computers & Electrical Engineering*, Elsevier, v. 40, n. 4, p. 1111–1122, 2014. 18

30 SASAKI, Y. et al. Daily sound recognition using pitch-cluster-maps for mobile robot audition. In: IEEE. *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. [S.l.], 2009. p. 2724–2729. 18

31 WANG, J.-C. et al. Robust environmental sound recognition for home automation. *Automation Science and Engineering, IEEE Transactions on*, IEEE, v. 5, n. 1, p. 25–31, 2008. 18

32 KIM, D.-S.; LEE, S.-Y.; KIL, R. M. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *Speech and Audio Processing, IEEE Transactions on*, IEEE, v. 7, n. 1, p. 55–69, 1999. 18

33 DAVE, N. Feature extraction methods lpc, plp and mfcc in speech recognition. *International Journal for Advance Research in Engineering and Technology*, v. 1, n. 6, p. 1–4, 2013. 18, 38

34 ANDÉN, J.; MALLAT, S. Multiscale scattering for audio classification. In: *ISMIR*. [S.l.: s.n.], 2011. p. 657–662. 18

35 FLETCHER, H.; STEINBERG, J. Articulation testing methods. *Bell System Technical Journal*, Wiley Online Library, v. 8, n. 4, p. 806–854, 1929. 18

36 GREENBERG, S. et al. *Speech processing in the auditory system*. [S.l.]: Springer, 2004. 18, 19, 20

37 RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, IEEE, v. 77, n. 2, p. 257–286, 1989. 20, 31, 47, 60

38 FURUI, S. On the use of hierarchical spectral dynamics in speech recognition. In: IEEE. *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. [S.l.], 1990. p. 789–792. 20, 22

39 DUMITRU, C. O.; GAVAT, I. A comparative study of feature extraction methods applied to continuous speech recognition in romanian language. In: IEEE. *Multimedia Signal Processing and Communications, 48th International Symposium ELMAR-2006 focused on*. [S.l.], 2006. p. 115–118. 20

40 RISSET, J.-C.; WESSEL, D. L. Exploration of timbre by analysis and synthesis. *The psychology of music*, Academic Press, v. 2, p. 151, 1999. 20

41 PLOMP, R. The ear as a frequency analyzer. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 36, n. 9, p. 1628–1636, 1964. 20

42 HELMHOLTZ, H. L. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. [S.l.]: Longmans, Green, 1912. 20, 21

43 GOLDSTEIN, J. Auditory spectral filtering and monaural phase perception. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 41, n. 2, p. 458–479, 1967. 20

44 MILLER, D. C. *The science of musical sounds*. [S.l.]: New York: The Macmillan Company, 1916. 20

45 STODOLSKY, D. S. The standardization of monaural phase. *Audio and Electroacoustics, IEEE Transactions on*, IEEE, v. 18, n. 3, p. 288–299, 1970. 20

46 MATHES, R.; MILLER, R. Phase effects in monaural perception. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 19, n. 5, p. 780–797, 1947. 20, 21

47 CRAIG, J. H.; JEFFRESS, L. A. Why helmholtz couldn't hear monaural phase effects. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 32, n. 7, p. 884–885, 1960. 20

48 BORWICK, J. *Loudspeaker and headphone handbook*. [S.l.]: CRC Press, 2012. 21

49 PLOMP, R.; STEENEKEN, H. Effect of phase on the timbre of complex tones. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 46, n. 2B, p. 409–421, 1969. 21

50 FURUI, S. Speaker-independent isolated word recognition based on emphasized spectral dynamics. In: IEEE. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86*. [S.l.], 1986. v. 11, p. 1991–1994. 21, 22, 43

51 GREY, J. M. *An exploration of musical timbre*. [S.l.]: Dept. of Music, Stanford University, 1975. 21, 22

52 JÄRVELÄINEN, H. et al. *Perception of attributes in real and synthetic string instrument sounds*. [S.l.]: Helsinki University of Technology, 2003. 21

53 GREY, J. M. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 61, n. 5, p. 1270–1277, 1977. 21

54 HANDEL, S. Timbre perception and auditory object identification. *Hearing*, Academic Press, San Diego, CA, p. 425–461, 1995. 21

55 FLETCHER, H. Auditory patterns. *Reviews of modern physics*, APS, v. 12, n. 1, p. 47, 1940. 22

56 MONTALVÃO, J.; ARAUJO, M. R. R. Is masking a relevant aspect lacking in mfcc? a speaker verification perspective. *Pattern Recognition Letters*, Elsevier, v. 33, n. 16, p. 2156–2165, 2012. 24

57 PELTONEN, V. et al. Computational auditory scene recognition. In: IEEE. *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. [S.l.], 2002. v. 2, p. II–1941. 24

58 KULKARNI, A.; IYER, D.; SRIDHARAN, S. R. Audio segmentation. In: CITESEER. *IEEE, International Conference on Data Mining, ICDM*. [S.l.], 2001. p. 105–110. 26

59 RYBACH, D. et al. Audio segmentation for speech recognition using segment features. In: IEEE. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. [S.l.], 2009. p. 4197–4200. 26

60 FOOTE, J. Automatic audio segmentation using a measure of audio novelty. In: IEEE. *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*. [S.l.], 2000. v. 1, p. 452–455. 26

61 GIANNAKOPOULOS, T.; PIKRAKIS, A. *Introduction to Audio Analysis: A MATLAB® Approach*. [S.l.]: Academic Press, 2014. 26, 36

62 BISHOP, C. M. *Pattern recognition and machine learning*. [S.l.]: springer, 2006. 26, 27, 28, 38, 39, 44

63 THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition*. Elsevier Science, 2006. (Pattern Recognition Series). ISBN 9780080513614. Disponível em: <https://books.google.com.br/books?id=gAGRCmp8Sp8C>. 26, 27, 28, 33, 44

64 GYÖRFI, L.; DEVROYE, L.; LUGOSI, G. *A probabilistic theory of pattern recognition*. [S.l.]: Springer-Verlag, 1996. 27

65 LI, X.; PARIZEAU, M.; PLAMONDON, R. Training hidden markov models with multiple observations-a combinatorial method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 22, n. 4, p. 371–377, 2000. 32

66 POVEY, D.; CHU, S. M.; VARADARAJAN, B. Universal background model based speech recognition. In: IEEE. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. [S.l.], 2008. p. 4561–4564. 33

67 CAMPBELL, W. M. et al. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In: IEEE. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. [S.l.], 2006. v. 1, p. I–I. 33

68 VACHER, M. et al. The sweet-home project: Audio technology in smart homes to improve well-being and reliance. In: IEEE. *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. [S.l.], 2011. p. 5291–5294. 34

69 MEDJAHED, H. et al. Human activities of daily living recognition using fuzzy logic for elderly home monitoring. In: IEEE. *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*. [S.l.], 2009. p. 2001–2006. 35

70 DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley & Sons, 2012. 36

71 HILL, M. The uncertainty principle for fourier transforms on the real line. 2013. 36

72 TZANETAKIS, G.; COOK, P. *Manipulation, analysis and retrieval systems for audio signals*. [S.l.]: Princeton University Princeton, NJ, USA, 2002. 36

73 BLÖMER, J.; BUJNA, K. Simple methods for initializing the em algorithm for gaussian mixture models. *arXiv preprint arXiv:1312.5946*, 2013. 39, 60

74 SHAUKAT, A. et al. Daily sound recognition for elderly people using ensemble methods. In: IEEE. *Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on*. [S.l.], 2014. p. 418–423. 41, 58

75 HOSSAN, M. A. et al. A novel approach for mfcc feature extraction. In: IEEE. *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*. [S.l.], 2010. p. 1–5. 42, 60

76 KUMAR, K.; KIM, C.; STERN, R. M. Delta-spectral cepstral coefficients for robust speech recognition. In: IEEE. *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. [S.l.], 2011. p. 4784–4787. 42, 43, 44

77 MUDA, L.; BEGAM, M.; ELAMVAZUTHI, I. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010. 43, 44, 46

78 GONZÁLEZ, D. R.; LARA, J. Calvo de. Speaker verification with shifted delta cepstral features: Its pseudo-prosodic behaviour. *Proc. I Iberian SLTech*, 2009. 43

79 ARORA, S. V. Effect of time derivatives of mfcc features on hmm based speech recognition system. *International Journal on Signal and Image Processing*, Association of Computer Electronics and Electrical Engineers (ACEEE), v. 4, n. 3, p. 50, 2013. 43

80 DEEMAGARN, A.; KAWTRAKUL, A. Thai connected digit speech recognition using hidden markov models. In: *9th Conference Speech and Computer*. [S.l.: s.n.], 2004. 46

81 STEVENS, S. S.; VOLKMANN, J.; NEWMAN, E. B. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 8, n. 3, p. 185–190, 1937. 54

82 BANDT, C.; POMPE, B. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, APS, v. 88, n. 17, p. 174102, 2002. 54

83 MEIGNIER, S.; BONASTRE, J.-F.; IGOUNET, S. E-hmm approach for learning and adapting sound models for speaker indexing. In: *2001: A Speaker Odyssey-The Speaker Recognition Workshop*. [S.l.: s.n.], 2001. 60

84 SJÖLANDER, K. An hmm-based system for automatic segmentation and alignment of speech. In: CITESEER. *Proceedings of Fonetik*. [S.l.], 2003. v. 2003, p. 93–96. 60

85 WANG, J.-C. et al. Environmental sound classification using hybrid svm/knn classifier and mpeg-7 audio low-level descriptor. In: IEEE. *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. [S.l.], 2006. p. 1731–1735. 60