Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Departamento de Engenharia Elétrica
Programa de Pós-Graduação em Engenharia Elétrica

Tese de Doutorado

Análise de Variações Acústicas Não Estacionárias e seu Efeito na Detecção de Múltiplas Emoções e Condições de Estresse

Vinícius Jefferson Dias Vieira

Prof. Francisco Marcos de Assis, Dr.
Profa. Rosângela Fernandes Coelho, Docteur ENST.

Orientadores

Campina Grande - PB, Brasil
Março - 2018

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Engenharia Elétrica

Análise de Variações Acústicas Não Estacionárias e seu Efeito na Detecção de Múltiplas Emoções e Condições de Estresse

Vinícius Jefferson Dias Vieira

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para obtenção do grau de Doutor em Ciências, no domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação.
Linha de Pesquisa: Eletrônica e Telecomunicações.

Prof. Francisco Marcos de Assis, Dr.
Profa. Rosângela Fernandes Coelho, Docteur ENST.
Orientadores

Campina Grande - PB
Março - 2018

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG(Bibliotecária Maria Antonia de Sousa CRB 15/398

V658a Vieira, Vinícius Jefferson Dias.

Análise de variações acústicas não estacionárias e seu efeito na detecção de múltiplas emoções e condições de estresse / Vinícius Jefferson Dias Vieira. – Campina Grande, 2018.

79 f. : il. color.

Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2018.

"Orientação: Prof. Dr. Francisco Marcos de Assis, Prof.ª Drª. Rosângela Fernandes Coelho".

Referências.

1. Atributo Acústico. 2. Decomposição Empírica de Modos. 3. Índice de Não Estacionariedade. 4. Reconhecimento de Emoções. I. Assis, Francisco Marcos de. II. Coelho, Rosângela Fernandes. III. Título.

CDU 007(043)

" ANÁLISE DE VARIAÇÕES ACÚSTICAS NÃO ESTACIONÁRIAS E SEU EFEITO NA DETECÇÃO DE MÚLTIPLAS EMOÇÕES E CONDIÇÕES DE ESTRESSE"

VINÍCIUS JEFFERSON DIAS VIEIRA

TESE APROVADA EM 02/03/2018

FRANCISCO MARCOS DE ASSIS, Dr., UFCG
Orientador(a)

ROSÂNGELA FERNANDES COELHO, Docteur, IME
Orientador(a)

BENEMAR ALENCAR DE SOUZA, D.Sc., UFCG
Examinador(a)

ALDEBARO BARRETO DA ROCHA KLAUTAU JÚNIOR, Dr., UFPA
Examinador(a)

BRUNO BARBOSA ALBERT, D.Sc., UFCG
Examinador(a)

WELLINGTON PINHEIRO DOS SANTOS, D.Sc., UFPE
Examinador(a)

CAMPINA GRANDE - PB

Aos meus pais, Verônica e Francisco.

Agradecimentos

Primeiramente, quero agradecer a Deus, Senhor da vida, por tudo que eu pude vivenciar até hoje, pelas pessoas que conheci, e por tudo que ainda está por vir. Aos meus pais, Verônica Regina e Francisco Vieira, por todo amor, educação, carinho e paciência para comigo, bem como à minha irmã Fernanda pela força.

Agradeço ao Professor Francisco Marcos de Assis, meu orientador, por me acolher na Universidade Federal de Campina Grande, pelo incentivo a continuar pesquisando e por sempre se disponibilizar a compartilhar de seus valiosos conhecimentos.

À Professora Rosângela Coelho, minha orientadora, por ter aceitado me orientar mesmo a quilômetros de distância, por abrir as portas do laboratório de processamento de sinais acústicos do IME, por toda a ajuda, incentivo e torcida. Agradeço também por sempre estar disposta a compartilhar seus valiosos conhecimentos.

Às pessoas com as quais criei um vínculo de amizade ao longo dos poucos anos de Academia até então. Entre elas, destaco as Professoras Silvana Costa, Suzete Correia, Taciana Souza (todas do IFPB) e os Professores Washington Costa (IFPB) e Leonardo Lopes (UFPB). Suas palavras de incentivo me ajudaram bastante.

Agradeço aos amigos de toda a vida que sempre torceram por mim. A todos os demais parentes. Ao meu primo Marcilho (que considero um irmão), que por inúmeras vezes me abriu as portas de sua casa para que eu pudesse estudar. Aos amigos que fiz no IQuanta, como Mikaelle, Juliana, Revson, Milena, Luiz Paulo e Micael, por todas as partilhas nos momentos de intervalo. Agradeço também aos colegas que fiz no IME que tanto me ajudaram, como o Guilherme Zucatelli e o Professor Leonardo Zão.

A todos os funcionários da COPELE, em especial à Ângela Matias, por toda sua dedicação ao programa e por sempre ter sido cordial para comigo, respondendo todas as dúvidas administrativas que eu pudesse ter ao longo do Doutorado, e por sua torcida pela minha conclusão.

Aos membros da Banca avaliadora desta Tese, pela disponibilidade e por suas valiosas contribuições.

E ao CNPq, pelo suporte financeiro.

"Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar.
Mas o mar seria menor se lhe faltasse uma gota."
(Madre Teresa de Calcutá)

Resumo

Nesta Tese, são estudados os efeitos das variações acústicas não estacionárias provocadas por estados emocionais e condições de estresse em sinais de voz. Ainda não há na literatura um atributo acústico puro para reconhecimento de emoções e estresse. Por meio do índice de não estacionariedade (*Index of Non-Stationarity* – INS), é observado que diferentes estados afetivos apresentam diferentes graus de não estacionariedade. Como forma de detectar tais variações, é empregada a decomposição empírica de modos (*Empirical Mode Decomposition* – EMD), que é uma técnica não linear adequada para sinais não estacionários. Com isso, a principal contribuição deste trabalho é a proposta do vetor HHHC (*Hilbert-Huang-Hurst Coefficients*) como um novo atributo acústico não linear para classificação multiestilo de estados emocionais e condições de estresse. O HHHC é um atributo da fonte de excitação que é baseado em decomposição adaptativa (EMD que enfatiza as variações acústicas afetivas) e estimação dos coeficientes de Hurst (que estão relacionados com a fonte de excitação glotal) em cada um dos modos da decomposição. Outra contribuição é a utilização do INS como informação adicional ao vetor HHHC (HHHC+INS). Para comprovar a robustez do atributo proposto em diferentes línguas e contextos de fala, são analisadas cinco bases de dados, sendo quatro delas no contexto de emoções e uma no contexto de condições de estresse. Como atributos acústicos comparativos ao HHHC, são utilizados o vetor de coeficientes de Hurst (pH), os coeficientes mel-cepstrais (*Mel-Frequency Cepstral Coefficients* – MFCC) e o atributo baseado no operador TEO (*Teager-Energy-Operator*). Outra importante contribuição desta Tese é a proposta dos modelos de misturas Gaussianas com integração α (α-*integrated Gaussian Mixture Models* – α-GMM) para representação e classificação dos estados afetivos. Seu desempenho é comparado com os seguintes métodos clássicos: Modelos de misturas Gaussianas (GMM), Modelos de Markov escondidos (*Hidden Markov Models* – HMM) e Máquinas de vetor de suporte (*Support Vector Machines* – SVM). Os resultados obtidos demonstram que o atributo proposto HHHC e sua fusão com o INS promovem taxas de acerto significativas em relação aos atributos comparativos. Além disso, o classificador α-GMM apresenta performance superior às técnicas comparativas em todos os cenários de bases acústicas.

Palavras-Chave: Atributo acústico, Decomposição empírica de modos, Índice de não estacionariedade, Reconhecimento de emoções.

Abstract

The goal of this work is to study the effects of non-stationary acoustic variations caused by emotional states and stress conditions. In the literature, there is still no pure acoustic attribute for emotion and stress recognition. By using the index of non-stationarity (INS), it is observed that different affective states have different degrees of non-stationarity. As for the detection of such variations, it is employed the empirical mode decomposition (EMD), which is a nonlinear technique that is suitable for non-stationary signals. Thus, the main contribution of this work is the proposal of the HHHC vector (Hilbert-Huang-Hurst Coefficients) as a new non-linear acoustic feature for the multistyle classification of emotional states and stress conditions. The HHHC is a vocal source feature that is based on adaptive decomposition (EMD, which emphasizes affective acoustic variations) and Hurst coefficients estimation (which are related to the glottal source excitation) in each decomposition mode. Another contribution is the use of INS as additional information to the HHHC vector (HHHC+INS). In order to analyze the robustness of the proposed acoustic feature in different languages and speaking contexts, it is considered five databases. Four of them in the context of emotions and one in the context of stress conditions. As baseline acoustic features to comparing with HHHC, it is used the vector of Hurst coefficients (pH), the Mel-Frequency Cepstral Coefficients (MFCC) and TEO (Teager-Energy-Operator)-based feature. Another important contribution of this Thesis is the proposal of α-integrated Gaussian Mixture Models (α-GMM) for the affective states representation and classification. Its performance is compared to competing classifiers: GMM, Hidden Markov Models (HMM) and Support Vector Machines (SVM). Results demonstrate that the proposed HHHC acoustic feature leads to significant classification improvement when compared to the baseline acoustic features. Also, the results show that α-GMM outperforms the competing classification methods in all acoustic databases scenarios.

Key-Words: Acoustic feature, Empirical mode decomposition, Index of nonstationarity, Emotion recognition.

Lista de Siglas e Abreviaturas

Alt. – Alto estresse
AMCC – *Amplitude Modulation Cepstral Coefficients*
DCT – *Discrete Cossine Transform*
Div. – Diversão
EMD – *Empirical Mode Decomposition*
EEMD – *Ensemble Empirical Mode Decomposition*
EMO-DB – *Berlin Database of Emotional Speech*
Fel. – Felicidade
FP – *Fourier Parameters*
GFCC – *Gammatone-Frequency Cepstral Coefficients*
GMM – *Gaussian Mixture Models*
Gri. – Grito
HHHC – *Hilbert-Huang-Hurst Coefficients*
HMM – *Hidden Markov Models*
IEMOCAP – *Interactive Emotional Dyadic Motion Capture*
IMF – *Intrinsic Mode Function*
INS – *Index of Non-Stationarity*
LLDs – *Low-Level Descriptors*
MFCC – *Mel-Frequency Cepstral Coefficients*
Med. – Médio estresse Neu. – Neutro
pH – vetor de coeficientes de Hurst
PSD – *Power Spectral Density*
Rai. – Raiva
RECOLA – *REmote COLaborative and Affective interactions*
SEMAINE – *Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression*
SUSAS – *Speech Under Simulated and Atual Stress*
SVM – *Support Vector Machine*
Ted. – Tédio TEO – *Teager Energy Operator*
Tri. – Tristeza

Lista de Símbolos

$x(t)$ – Sinal acústico
$X[K]$ – Transformada de Fourier de $x(t)$
$A[k]$ – Amplitude de $X[K]$
$\theta[k]$ – Fase de $X[K]$
IMF$_k$ – k-ésima função intrínseca de modo
$S_{t,f}$ – espectograma de $x(t)$
c_n – distância KL entre os espectogramas
Θ – medida de invariância no cálculo do INS

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Quadros

CAPÍTULO 1

Introdução

Estados afetivos estão presentes no cotidiano dos seres humanos, influenciando a cognição, a percepção, o aprendizado e a comunicação. Um evento inesperado, por exemplo, pode ser motivado por algo que gere a sensação de felicidade. Por outro lado, uma surpresa negativa pode ocorrer, levando o indivíduo a sensações como medo e estresse, o que afetaria, entre outras coisas, tomadas de decisão de curto prazo.

A voz é considerada como o meio mais natural de comunicação do ser humano. A ideia de que ela contém informações sobre o locutor, não diz respeito apenas à sua identidade ou ao conteúdo linguístico (ou sentido literal) da fala. Recentemente, aspectos prosódicos têm sido observados na emissão sonora, fazendo com que começassem a ser estudadas variações acústicas provocadas por diferentes estados afetivos presentes na produção da voz [4–7]. Dessa forma, pesquisas neste contexto estão voltadas não apenas para questões como "quem fala?" ou "o que é falado?", mas também procura saber "como é falado", ou seja, tenta identificar as emoções contidas na emissão sonora.

Classificação de emoções e de condições de estresse tem recebido atenção nos últimos anos [4–7]. Diversas aplicações biométricas relacionadas à análise de variações acústicas afetivas podem ser citadas, tais como: sistemas de segurança, sistemas de bordo de veículos (nos quais a informação do estado mental do condutor é útil para a sua própria segurança), sistemas de tradução automática em conversações (nas quais é útil que se entenda o estado emocional de ambas as partes) sistemas de *call centers* e sistemas de iteração humano-robô [4, 8, 9]. Por isso, um sistema robusto de iteração homem-máquina é necessário para extrair características significantes e detectar de forma precisa emoções e condições de estresse [5].

Nas iterações sociais, há uma grande variedade de estados emocionais (a exemplo de Raiva, Felicidade e Tristeza) [10]. De acordo com Ekman [11], há certas emoções que são naturalmente reconhecidas pelos seres humanos. Embora haja essa universalidade na discriminação de estados afetivos, a sua representação em um sistema homem-máquina ainda é um grande desafio. Uma "impressão acústica afetiva" é fundamental para um poderoso sistema de reconhecimento. Assim, um importante desafio é definir um atributo que caracterize

diferentes estados emocionais e condições de estresse [4, 9]. Na literatura, ainda não há um consenso a respeito de um atributo acústico afetivo para esta tarefa. Dessa forma, a definição de um atributo que represente de forma significativa informações relacionadas ao comportamento fisiológico dos estados afetivos é uma busca crucial.

A análise e classificação de estados afetivos para a definição de um sistema que extraia uma "impressão acústica afetiva" e que seja independente de locutor e de texto deve levar em consideração questões como:

- A variabilidade acústica introduzida pela existência de diferentes sentenças, locutores e estilos de fala;
- Diferentes idiomas;
- Diferentes tarefas, a exemplo de reconhecimento de emoções e reconhecimento de condições de estresse;
- Ponto de vista dos efeitos das variações afetivas (fonte de excitação glotal *versus* trato vocal, por exemplo).

Geralmente, o processo de classificação engloba três procedimentos, que devem ser levados em consideração após a aquisição do sinal de voz: pré-processamento, extração da matriz de atributos e classificação da emoção ou da condição de estresse. A etapa de pré-processamento é a primeira coisa a ser realizada após a aquisição do sinal. Por meio dela é que se define, por exemplo, qual a escala da análise (tamanho dos quadros de voz) e em que domínio o sinal será analisado (tempo ou frequência). Na extração da matriz de atributos é necessário que sejam empregadas medidas que possam detectar a variação acústica afetiva presente no sinal. Para tanto, as medidas são extraídas de acordo com a etapa de pré-processamento (por exemplo, o tamanho da matriz de atributos depende da quantidade de quadros de voz). A classificação da emoção/condição de estresse consiste na resposta do sistema ao sinal de entrada. Nesta etapa, a matriz de atributos obtida do sinal de entrada na interface é comparada com padrões armazenados para diferentes estados afetivos. Assim, uma classificação é realizada e a decisão é tomada de acordo com o padrão acústico de variação afetiva que mais se aproxime do que foi observado no sinal de teste.

Atividade muscular, diâmetro da pupila, pressão sanguínea e batimentos cardíacos são alguns aspectos fisiológicos levados em consideração na análise de variações acústicas afetivas. Emoções como Raiva e Felicidade, por exemplo, acarretam aumento na pressão sanguínea, nos batimentos cardíacos e alteração nos movimentos respiratórios [4, 5]. No que diz respeito à produção vocal, alterações deste tipo provocam mudanças no fluxo glotal [12], tendo como consequência alterações na densidade espectral de potência (*Power Spectral Density* – PSD) do sinal resultante. Sem variação acústica emocional (estado Neutro), sinais de voz apresentam uma densidade espectral de potência com queda, em média, de 12dB/oitava. A variação de

3dB/oitava (para mais ou para menos) do estado Neutro define os estados emocionais como sendo de alta e de baixa ativação (9dB/oitava e 15dB/oitava, respectivamente). Sinais de voz de alta ativação apresentam maior variação e, portanto, verifica-se concentração superior de energia nas altas frequências. Para a ocorrência de estados emocionais de baixa ativação, o sinal apresenta pouca mudança e desta forma exibe maior concentração de energia nas baixas frequências [13, 14].

Os fatores fisiológicos supracitados ainda induzem a estruturas de fluxo de ar dinâmicas e não lineares no sinal de voz [12]. O atributo baseado no operador de energia Teager (*Teager-Energy-Operator* – TEO) foi proposto considerando este conceito no contexto da classificação de condições de estresse. Considerando os efeitos afetivos na fonte de excitação glotal, atributos voltados a este aspecto podem ser empregados no sentido de serem menos dependentes do conteúdo linguístico em comparação a atributos do trato vocal [15]. Em [7], o atributo pH da fonte de excitação foi proposto para classificação de emoções e estresse. Os autores mostraram que o atributo baseado em TEO pode não ser adequado para identificação de emoções. Tanto TEO quanto pH não levam em consideração alguns efeitos da produção não linear da voz sob efeito de estados afetivos, tais como a não estacionariedade das variações acústicas e seu comportamento dinâmico. Estes fatores são importantes para serem explorados por um atributo acústico.

Nesta Tese é proposto um novo vetor de atributos acústicos não linear baseado nas variações acústicas ou nos efeitos de emoções e estresse na fonte de excitação glotal. A ideia básica é o fato de que estados afetivos são variações acústicas não estacionárias introduzidas na produção da voz. A abordagem baseada em decomposição empírica de modos (*Empirical Mode Decomposition* – EMD) [16] é aplicada na tarefa de detecção dessas variações. Coeficientes de Hurst [17] são adotados para caracterizar as componentes da fonte de excitação que são enfatizadas pela decomposição dos sinais. O vetor de atributos HHHC (*Hilbert-Huang-Hurst Coefficients)* é então formado em uma extração quadro a quadro de cada função resultante da decomposição. O índice de não estacionariedade (*Index of Non-Stationarity* – INS) [18] também é proposto como informação adicional ao vetor HHHC. Esta medida descreve dinamicamente o comportamento não estacionário das amostras de voz. Além destas contribuições, o modelo α-GMM [19] é proposto para classificar os diferentes estados emocionais e condições de estresse. Na análise comparativa, utiliza-se classificadores clássicos estocásticos (GMM [20] e HMM [21]) e de aprendizado de máquina (SVM [22]). Os experimentos, realizados em cinco bases acústicas, mostram a efetividade do novo atributo acústico da fonte de excitação em diferentes idiomas. Os resultados demonstram que HHHC com um vetor de dimensão 6 que alcança robustez de um atributo acústico puro de emoções e estresse. A informação do INS junto ao vetor HHHC leva a acréscimos nas taxas de acerto da classificação. Além disso, o atributo proposto α-GMM obteve desempenho superior aos classificadores comparativos.

1.1 – Estado da Arte

Nesta Seção, são apresentados trabalhos relevantes relacionados à análise e classificação de variações acústicas afetivas. Com o crescente surgimento de estudos relacionados a classificação/reconhecimento de estados emocionais e de condições de estresse, o principal desafio é encontrar a melhor característica ou o melhor conjunto de características que seja o mais apropriado possível para esta tarefa [4, 9]. Na literatura, ainda não há um consenso a respeito de um atributo acústico que seja relevante na detecção e classificação de estados afetivos.

Algumas abordagens têm utilizado atributos prosódicos, tais como frequência fundamental ($F0$) e energia [23, 24]. Por outro lado, o atributo acústico mais comumente encontrado na literatura é baseado em coeficientes mel cepstrais (*Mel-Frequency Cepstral Coefficients* – MFCC) [6, 7, 25]. Vários estudos utilizam os coeficientes MFCC para propósitos de comparação de desempenho. Em [26], atributos de qualidade vocal (*jitter* and *shimmer*) foram empregados na classificação de condições de estresse. O atributo proposto obteve uma performance de 3 pontos percentuais (p.p.) a mais que o clássico MFCC.

Além do MFCC, outros atributos baseados em características espectrais têm sido propostos em tarefas de computação afetiva. Entre esses atributos estão medidas baseadas em modulação [25, 27]. Em [25], o vetor de coeficientes cepstrais de modulação em amplitude (*Amplitude Modulation Cepstral Coefficients* – AMCC) foi proposto por meio da utilização de um banco de filtros *gammatone*, formando um vetor de dimensão 39. Os resultados demonstraram que o atributo AMCC obteve acurácia 3,5 p.p. acima do desempenho do MFCC na classificação de estados emocionais. Em [28], coeficientes da transformada de Fourier (*Fourier Parameters* – FP) foram aplicados para classificação de emoções. Os autores experimentaram vários tamanhos do vetor com o objetivo de aprimorar as taxas de acerto. Um procedimento de seleção de características foi realizado e então um vetor de 120 coeficientes atingiu uma taxa de acerto 16,2 p.p. superior do que foi obtido com MFCC. No entanto, por se tratarem de dois atributos espectrais, esta diferença de acurácia pode ser devido à alta dimensionalidade do vetor FP em relação ao MFCC.

Uma outra questão no contexto de classificação de variações acústicas afetivas é o tamanho do atributo ou da matriz de atributos. Vários estudos têm focado em combinar diferentes atributos a fim de testar tanto o conjunto de características quanto o esquema de classificação [29]. Em alguns casos, mais de 1000 medidas para compor um vetor de atributos [28, 30]. Por exemplo, na proposta apresentada em [28], o vetor original contém 1800 componentes antes do procedimento de seleção de características. Na literatura, esta combinação de atributos acústicos tem sido chamada de conjunto de descritores de baixo nível (*Low-Level Descriptors* – LLDs) [9, 31]. Geralmente estes descritores são acompanhados dos chamados "funcionais", que são medidas derivadas dos LLDs, a exemplo de estatísticas como média e mediana. Em [31] foram empregados 88 parâmetros no conjunto

de características, incluindo LLDs e vários funcionais. O uso de funcionais pode ser uma questão aberta na classificação das variações acústicas afetivas, no que diz respeito à sua interpretação física na formação dessas variações acústicas. Por exemplo, embora os LLDs empregados em [31] estejam relacionados ao mecanismo de produção/percepção da voz, eles precisam de várias informações adicionais (os funcionais) para obter uma melhora nas taxas de acerto. Ainda, mesmo que estes tipos de procedimento proporcionem bons resultados em termos de acurácia, eles podem encontrar problemas em situações de tempo real. Para sistemas que necessitam de resposta de curto prazo, robustez significa matriz de atributos relativamente pequena e que proporcione bons resultados de classificação [9]. Neste contexto, a construção de um atributo de pequena dimensionalidade que extraia informação significante relacionada à fisiologia dos estados afetivos é uma escolha interessante.

Em [12], os autores propuseram atributos baseados no operador de energia TEO (*Teager Energy Operator*) para analisar diferentes condições de estresse. Eles demosntraram que aspectos fisiológicos refletem em variações acústicas afetivas. O atributo TEO baseia-se em uma abordagem não linear da produção da voz, apresentada nos estudos de Teager [32] e Kaiser [33]. Em [7], um vetor de 12 dimensões com expoentes de Hurst (vetor pH) foi proposto para classificação de variações afetivas, em que atingiu uma acurácia 6,8 p.p a mais que MFCC e 17,7 p.p. a mais que o atributo baseado em TEO. Os autores mostraram que o expoente de Hurst está relacionado com os estados emocionais.

1.2 – Objetivos

Os principais objetivos deste trabalho são:

- Analisar variações acústicas não estacionárias presentes nos sinais de voz a fim de detectar estados emocionais e condições de estresse. Para isto, é utilizado o método EMD para decompor os sinais acústicos. Além disso, é empregado o INS a fim de que seja observado o grau de não estacionariedade dos estados afetivos;

- Propor um novo atributo acústico que seja capaz de extrair informações ou características relevantes de cada estado emocional presente nas variações acústicas não estacionárias, ou seja, um atributo que seja uma impressão acústica dos estados afetivos. Para isto, é utilizado o expoente de Hurst para capturar informação não linear de cada função resultante da decomposição dos sinais. Assim, as variações acústicas são destacadas e associadas a cada estado afetivo;

- Analisar o desempenho do atributo acústico proposto em diferentes bases acústicas de emoção em diferentes idiomas e contextos de gravação. Além disso, analisar a robustez do atributo acústico no contexto de diferentes condições de estresse;

- Propor o α-GMM como método de classificação das variações acústicas afetivas. Para análise comparativa, utilizar classificadores clássicos, tais como HMM e SVM, a fim de se investigar as taxas de acerto do atributo;

- Comparar o desempenho do atributo acústico proposto com outros atributos apresentados na literatura (pH, MFCC e TEO), no procedimento de classificação. Ainda, realizar a fusão do novo atributo com os demais atributos considerados para investigar quanto de melhora é proporcionada pelo novo atributo aos clássicos;

1.3 – Resultados Obtidos

Os principais resultados e contribuições alcançados nesta Tese são os seguintes:

- Com a análise baseada em EMD, foi possível enfatizar as variações acústicas afetivas presentes nos sinais de voz. Ainda neste contexto, foi observado por meio do INS que os estados afetivos possuem diferentes graus de não estacionariedade;

- Os resultados com o novo atributo acústico, HHHC, demonstraram superioridade aos atributos comparativos utilizados nesta pesquisa. Desta forma, foi possível demonstrar que o HHHC e sua fusão com o INS é uma impressão acústica afetiva de estados emocionais e de condições de estresse. As taxas de acerto com HHHC foram superiores com todos os classificadores utilizados;

- A partir da utilização de diferentes classificadores, foi observado que aqueles de abordagem estocástica apresentaram melhor desempenho. Adicionalmente, o α-GMM obteve as maiores taxas de acerto em relação a GMM, HMM e SVM.

- Os resultados com HHHC foram superiores aos atributos comparativos considerando todas as bases acústicas empregas nesta Tese. Ainda foi observado que o HHHC agrega valor às taxas de acerto de cada atributo comparativo. Também para todas as bases acústicas consideradas, o α-GMM apresentou-se como classificador mais apropriado nesta tarefa em relação aos classificadores clássicos.

1.4 – Organização da Tese

Além deste Capítulo introdutório, esta Tese está organizada da seguinte maneira:

- **Capítulo 2:** Neste Capítulo, são apresentadas as principais teorias relacionadas à definição de estados afetivos. A partir de um ponto de vista de que os estados emocionais e as condições de estresse são variações acústicas não estacionárias introduzidas na voz, são apresentados os principais métodos empregados nesta Tese para esta tarefa: a EMD e o INS;

- **Capítulo 3:** Um novo atributo acústico para classificação de variações acústicas afetivas não estacionárias é proposto neste Capítulo: o vetor HHHC. Por meio das análises realizadas, é apresentada a configuração mais apropriada para este atributo, bem como sua fusão com o INS de forma a acrescentar informação sobre o comportamento da não estacionariedade das variações afetivas na matriz de atributos;

- **Capítulo 4:** Neste Capítulo, é apresentada a metodologia empregada nos experimentos, bem como as técnicas de classificação utilizadas, as bases acústicas analisadas e os atributos comparados com o HHHC. No contexto da classificação, é proposto o α-GMM para classificação de estados afetivos. Ainda, são apresentados os resultados desta etapa de classificação das variações acústicas afetivas;

- **Capítulo 5:** Finalmente, são apresentadas neste Capítulo as considerações finais e as contribuições desta Tese. Ainda, são apresentadas sugestões para trabalhos futuros.

CAPÍTULO 2

Variações Acústicas Afetivas Não Estacionárias

Neste Capítulo, são apresentados conceitos e definições relacionados a variações acústicas afetivas e os métodos empregados para análise dessas variações. Estados afetivos estão presentes no contexto diário dos seres humanos sob diferentes aspectos. O estado físico e o ambiente de iteração com outras pessoas, por exemplo, são fatores que podem ser importantes na condição emocional do sujeito. Assim, as variações afetivas influenciam em comportamentos fisiológicos, como a produção da voz. Emoções ou condições de estresse, então, constituem-se de variações acústicas não estacionárias introduzidas na fala. Dessa perspectiva, é realizada nesta Tese uma análise dessas variações de modo a definir um padrão acústico que represente cada estado emocional e cada condição de estresse. Para a análise, são propostos métodos, índice de não estacionariedade (*Index of non Stationarity* – INS) para quantificar o grau de não estacionariedade das variações acústicas e a decomposição empírica de modos (*Empirical Mode Decomposition* – EMD) para a tarefa de detecção destas variações.

2.1 – Conceitos Relacionados a Estados Afetivos

Estados afetivos, de uma maneira geral, descrevem sentimentos subjetivos em curtos períodos de tempo que retratam as experiências humanas com eventos, pessoas ou objetos [5, 34]. Diferentes fenômenos, neste contexto, representam estados afetivos, tais como emoções, humor, condições de estresse, posições interpessoais e traços de personalidade [35]. Raiva e Tristeza, por exemplo, são aspectos considerados como emoções, enquanto que "estar distante" ou "frio(a)" caracterizam-se como posições interpessoais. Por outro lado, condições de estresse podem incluir traços de personalidade, humor e emoções, envolvendo situações de médio e alto estresse, inclusive grito [36].

Dentre as representações dos estados afetivos, as emoções constituem-se do principal objeto de estudo nos últimos anos [4–9]. Diversas áreas do conhecimento envolvem o estudo das

emoções nas mais diferentes aplicações [35]:

- **Filosofia**: estudo de ética, artes e música;
- **Ciências Sociais**: sociologia e antropologia;
- **Ciências Humanas**: psicologia, psiquiatria, neurociência e linguística;
- **Política**: influência na decisão de voto;
- **Economia**: influência na decisão de compra;
- **Eologia**: comportamento animal e biologia evolucionária;
- **Engenharia/Computação**: tarefas de análise, detecção, reconhecimento.

Uma vez que o estado emocional do ser humano é resultado de experiências altamente subjetivas, é difícil encontrar definições universais.

2.1.1 – Teorias Relacionadas a Emoções

• Do Ponto de Vista Filosófico

As teorias relacionadas a emoções existem desde a Grécia antiga e têm sido debatidas até os dias atuais. Dentre as concepções mais importantes, a primeira que pode ser citada é a do filósofo Platão (427–347 a.C.), a qual sugeria que a emoção era um de três pilares que formavam a estrutura da alma, juntamente com cognição e motivação. Nesta definição, estes três pilares constituem de áreas opostas. Aristóteles (384–322 a.C.), por sua vez, argumentava que estes níveis de funcionamento psicológico são interligados [35]. O próprio Aristóteles já definia emoções específicas, como a Raiva, que para o filósofo trata-se de um desejo acompanhado de dor e de vingança percebida. Neste contexto, dor e prazer não se trata como emoções, e sim como sensações [37]. Outro filósofo, Descartes (1596–1650 d.C.), propôs que razão e emoção não estão conectados, o que foi refutado pelo neurocientista Damásio em 1994 d.C. [38].

• Do Ponto de Vista Biológico

Em um contexto biológico, Darwin (1809–1882 d.C.) lançou um trabalho sobre a expressão de emoções no homem e nos animais [39]. Ele apontou que expressões faciais e movimentos corporais são padrões de ação que ligam as emoções com o processo de seleção natural. Neste contexto, emoções representavam um aspecto de sobrevivência, o que seria uma característica de todo ser humano. Dessa forma, na visão de Darwin, emoções são fatores independentes de cultura. No caminho de Darwin, outros pesquisadores aprofundaram o estudo sobre a universalidade das emoções. Nesses estudos, compreende-se que existem emoções básicas que, quando combinadas, geram um espectro de estados emocionais. Ekman [11], por exemplo,

definiu um grupo de seis emoções básicas, conhecido na literatura como "Big-Six". Estas emoções são: Raiva, Felicidade, Surpresa, Desgosto (ou Nojo), Medo e Tristeza. Outro estudo desenvolvido neste sentido foi o de Plutchik [1, 40], que definiu oito emoções básicas: Raiva, Felicidade, Surpresa, Desgosto (ou Nojo), Tristeza, Curiosidade e Aceitação. Na Figura 2.1 está representada a chamada "Roda de Plutchik".

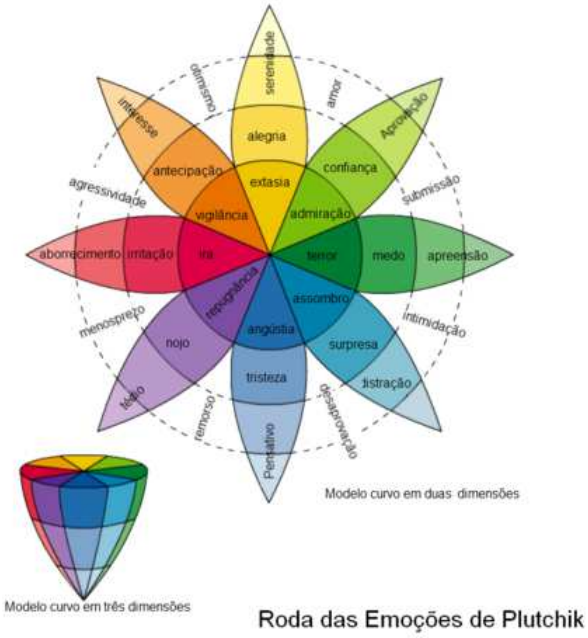


Figura 2.1 – Roda das Emoções de Plutchik (Adaptado de [1]).

A categorização dos estados emocionais em emoções básicas também é conhecido na literatura como abordagem discreta das emoções [2, 41]. Por outro lado, existe a abordagem contínua, que categoriza as emoções em eixos de N dimensões. O modelo mais conhecido, neste sentido, é o modelo tridimensional de Schlosberg [42]. Cada estado emocional pode ser definido como uma combinação linear dos eixos Ativação (ou Excitação), Valência (ou Avaliação) e Potência (ou Poder). Ativação mede o grau de excitação do indivíduo em expressar a emoção. Valência quão positiva ou negativa é a emoção. Potência diz respeito à força da emoção. Na Figura 2.2 é apresentado o sistema de eixos de Schlosberg com a distribuição das emoções discretas ao longo das três dimensões.

Outra teoria que tem sido bastante considerada é de que os estados emocionais constituem-se de reações fisiológicas a estímulos, formulada por William James e Carl Lange [43]. Enquanto os estudos de Darwin buscavam interpretar como as emoções são expressas, a teoria de James-Lange busca explicar a natureza da experiência emocional. Nesta abordagem, as emoções são sentimentos sobre alterações fisiológicas tais como alteração do batimento cardíaco, tensão muscular e transpiração, que são resposta da experiência do indivíduo com o mundo.

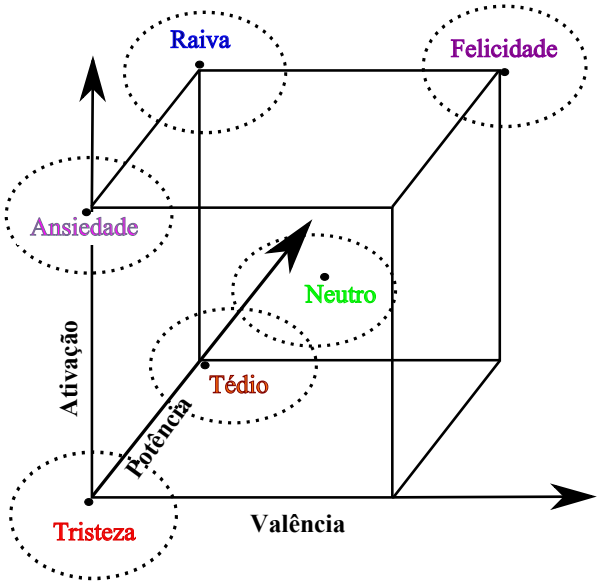


Figura 2.2 – Eixos das Emoções de Schlosberg (Adaptado de [2]).

• Do Ponto de Vista de Processamento de Sinais

A teorias supracitadas foram principalmente desenvolvidas a partir de estudos baseados nas expressões faciais e movimentos corporais. Porém, como as variações emocionais tem influência na fisiologia humana, elas podem ser analisadas a partir de sinais biológicos, tais como eletrocardiograma (ECG) e Eletroencefalograma (EEG) [44, 45]. A voz também é um sinal biológico, o qual é produzido a partir da geração de um fluxo de ar nos pulmões que passa pela laringe e pelas cavidades do trato vocal até a saída. Dessa forma, qualquer alteração fisiológica resultará, na fala, em um sinal acústico diferente do que seria seu estado Neutro.

No contexto da produção vocal, alterações fisiológicas provocam mudanças no fluxo glotal [12], tendo como consequência alterações na densidade espectral de potência do sinal de voz resultante. Sinais de voz que não apresentam emoção (estado Neutro), apresentam uma densidade espectral de potência com queda, em média, de 12dB/oitava. A variação de 3dB/oitava (para mais ou para menos) do estado Neutro define os estados emocionais como sendo de alta e de baixa ativação (9dB/oitava e 15dB/oitava, respectivamente). Sinais de voz de alta ativação apresentam maior variação e, portanto, verifica-se concentração superior de energia nas altas frequências. Para a ocorrência de estados emocionais de baixa ativação, o sinal apresenta pouca mudança e desta forma exibe maior concentração de energia nas baixas frequências [13, 14].

A partir dos estudos de Teager [32] e Kaiser [33], nas últimas décadas técnicas de análise não linear tem sido empregadas em aplicações de processamento de sinais de voz, levando em consideração fatores que indicam a presença de não linearidades no sistema de produção da voz. Entre esses fatores estão a variação temporal da forma do trato vocal, as ressonâncias associadas à sua fisiologia, as perdas devido ao atrito viscoso nas paredes internas do trato

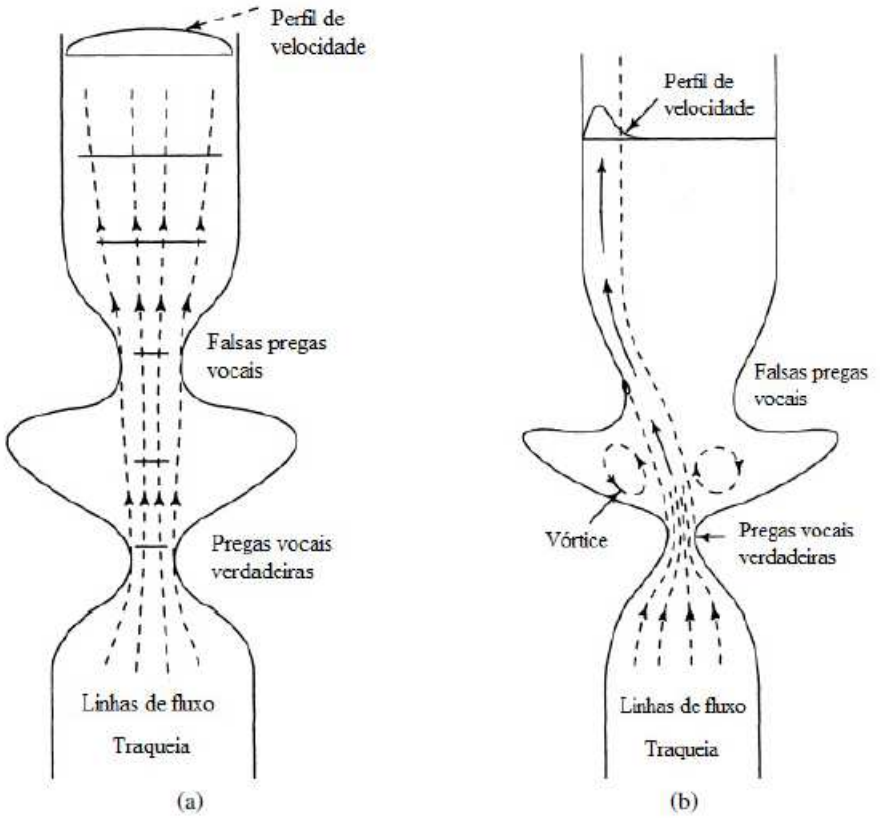


Figura 2.3 – (a) Interpretação clássica da propagação do som através do sistema vocal. (b) Interpretação da dinâmica não linear de fluidos para a propagação do som ao longo do sistema vocal [3].

vocal, a suavidade dessas paredes internas, a radiação do som nos lábios, o acoplamento nasal e a flexibilidade (comportamento dinâmico) associada à vibração das pregas vocais [46]. Esses fatores causam os chamados vórtices na onda sonora, como mostra a Figura 2.3. Indícios que essas não linearidades são influenciadas por estados afetivos foram descritos em [12].

Variações acústicas em sinais de voz podem ser compreendidas como sendo processos adicionados ao sinal em seu estado Neutro. Por exemplo, ruídos são variações acústicas que são introduzidas no sinal de voz ao longo de um sistema de comunicações. Neste sentido, os estados emocionais são variações acústicas introduzidas na voz durante o seu processo de produção e são eventos ocasionados em curtos períodos de tempo [9]. A partir da produção não linear da voz, pode ser investigada a natureza não estacionária das variações acústicas afetivas. Este ponto de vista não tem sido explorado na literatura, pelo que tem sido pesquisado, sendo assim relevante para as contribuições apresentadas neste trabalho de Tese.

2.2 – Índice de Não Estacionariedade

A estacionariedade é um aspecto relevante em muitas aplicações de processamento de sinais [18]. Diferentes métodos têm sido propostos na literatura como forma de medir a não estacionariedade de sinais [18, 47–50]. Nesta Tese, foi escolhido o método que propõe a medida

do índice de não estacionariedade (*index of nonstationarity* – INS), como forma de investigar o comportamento das variações acústicas afetivas.

Por definição, o INS é um método tempo-frequência para estimação objetiva do grau de não estacionariedade de sinais [18]. Esta avaliação é realizada a partir da comparação das componentes espectrais do sinal original e de referenciais espectrais estacionários (*surrogates*).

Dada a transformada discreta de Fourier, $X[k]$, do sinal $x(t)$, escrita em termos de sua magnitude, $A[k]$, e sua fase, $\phi[k]$,

$$X[k] = A[k] \exp(i\,\phi[k])\,, \tag{2.1}$$

em que i é a unidade imaginária. Um referencial "estacionário", $\tilde{x}(t)$, do sinal original pode ser obtido aplicando-se a transformada inversa de Fourier de $\tilde{X}[k]$:

$$\tilde{X}[k] = A[k] \exp(i\,\psi[k])\,, \tag{2.2}$$

em que $\psi[k]$ é uma sequência aleatória com amostras independentes e uniformemente distribuídas no intervalo $[-\pi, \pi]$. Um conjunto de sinais referenciais ($\{\tilde{x}_j(t), j = 1, ..., J\}$) é obtido repetindo-se este procedimento para J sequências aleatórias $\psi[k]$.

Para que o sinal $x(t)$ seja comparado com os seus referenciais estacionários obtidos, primeiramente é calculado o seu espectrograma por meio de um janelamento multi-ortogonal (*multitaper*), definido por

$$S_{x,K}(t,f) = \frac{1}{K}\sum_{k=1}^{K} S_x^{(h_k)}(t,f) \tag{2.3}$$

O janelamento é realizado por meio de K funções de Hermite (h_k) definidas em janelas de tempo curto por

$$h_k(t) = g(t)H_k(t)/\sqrt{\pi^{1/2}2^k k!}, \tag{2.4}$$

em que $g(t) = \exp\{-t^2/2\}$ e $\{H_k(t), t \in \mathbb{N}\}$ representa polinômios de Hermite, que obedecem a recursão

$$H_k(t) = 2tH_{k-1}(t) - 2(k-2)H_{k-2}(t), \quad k \geq 2, \tag{2.5}$$

com inicialização $H_0(t) = 1$ e $H_1(t) = 2t$.

Na Equação 2.3, $\{S_x^{(h_k)}, k = 1, 2, ..., K\}$ são os K espectrogramas calculados com as K primeiras funções de Hermite como janelas de tempo curto, $h_k(t)$:

$$S_x^{(h_k)}(t,f) = \left| \int x(s)h_k(s-t)e^{-i2\pi fs}ds \right|^2. \tag{2.6}$$

Assim, se os espectrogramas (Equações 2.3, 2.4 e 2.5) são avaliados em diversos pontos

$\{t_1, t_2, ..., t_{N_p}\}$, a média dos espetrogramas de $x(t)$ é construída segundo

$$\langle S_{x,K}(t_n, f)\rangle_n := \frac{1}{N_p}\sum_{n=1}^{N_p} S_{x,K}(t_n, f). \tag{2.7}$$

Essa média dos espectrogramas (Equação 2.7) é comparada com os próprios espectrogramas obtidos em cada um dos pontos $\{t_1, t_2, ..., t_{N_p}\}$, por meio de uma métrica de distância conhecida como Distância de Kullback-Leibler (D_{KL}) simétrica [51], de acordo com

$$\{c_n^{(x)} := D_{KL}(S_{x,K}(t_n, .), \langle S_{x,K}(t_n, .)\rangle_n), \quad n = 1, ..., N\}. \tag{2.8}$$

A distância $D_{KL}(\cdot, \cdot)$ para duas funções $G(f)$ e $B(f)$ é dada por

$$D_{KL}(G, B) := \int (G(f) - B(f))\log\frac{G(f)}{B(f)}df. \tag{2.9}$$

O conjunto de valores das distâncias D_{KL}, obtidos de todos os referenciais estacionários, é definido da seguinte forma:

$$\{c_n^{(\tilde{x}_j)} := D_{KL}(S_{\tilde{x}_j,K}(t_n, .), \langle S_{\tilde{x}_j,K}(t_n, .)\rangle_n), \quad n = 1, ..., N\}, \quad j = 1, 2, ..., J. \tag{2.10}$$

O INS é então definido como a razão entre a variância das distâncias observadas do sinal em análise e a média das variâncias obtidas dos sinais referenciais. Ou seja,

$$\text{INS} := \sqrt{\frac{\Theta_1}{\langle\Theta_0(j)\rangle_j}}, \tag{2.11}$$

em que:

$$\begin{cases} \Theta_0(j) = \text{Var}\left(c_n^{(\tilde{x}_j)}\right)_{n=1,...,N.}, \quad j = 1, ..., J. \\ \Theta_1 = \text{Var}\left(c_n^{(x)}\right)_{n=1,...,N.} \end{cases} \tag{2.12}$$

Caso o INS seja maior que um certo limiar γ, definido para uma precisão de 95%, o sinal é considerado não estacionário [18]. Ou seja,

$$\text{INS}\begin{cases} \leq \gamma & , x(t) \text{ é estacionário;} \\ > \gamma & , x(t) \text{ não é estacionário.} \end{cases} \tag{2.13}$$

Na Figura 2.4 são apresentados exemplos de INS obtido a partir de sinais de voz no estado Neutro e em dois diferentes estados emocionais: Raiva e Tristeza. A escala temporal T_h/T representa a razão entre o tamanho da janela de análise espectral de tempo curto (T_h) e a duração total (T =800 ms) do sinal. Note que o INS para os estados emocionais (linha vermelha) é maior que o limiar adotado no teste de não estacionariedade (linha verde). Embora os sinais de

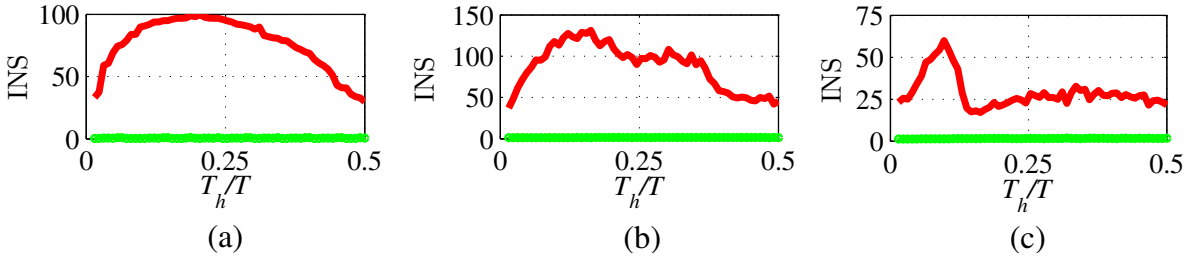


Figura 2.4 – INS calculado das seguintes emoções: (a) Neutro, (b) Raiva e (c) Tristeza.

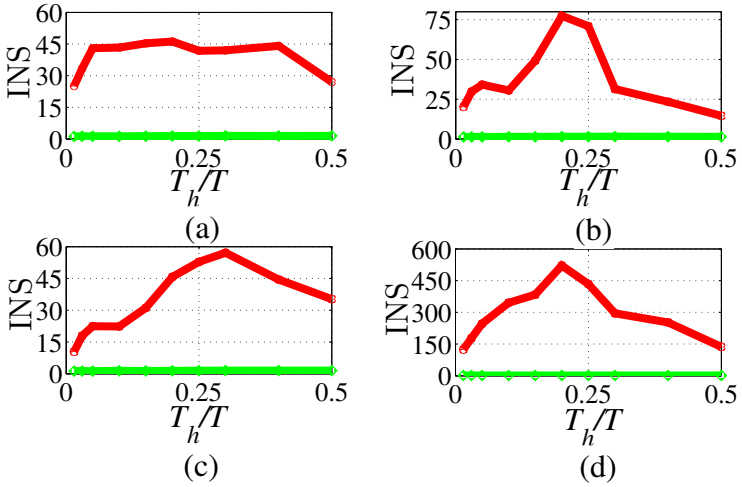


Figura 2.5 – INS calculado das seguintes condições de estresse: (a) Neutro, (b) Médio estresse, (c) Alto estresse, e (d) Grito.

voz sejam processos não estacionários, o grau de não estacionariedade difere de uma emoção para outra. Enquanto o estado Neutro tem valores de INS no intervalo [50,100] para a maioria das escalas de observação, o INS para Tristeza atinge um valor máximo de 60. Por outro lado, Raiva atinge INS maior que 100 em algumas das escalas de tempo.

Como exemplo da análise do INS no caso de condições de estresse, é mostrada a Figura 2.5 com três condições de estresse além do estado Neutro: Médio, Alto e Grito. Para o estado Neutro, o INS é aproximadamente constante em quase todo o intervalo de observação. As condições de estresse apresentam diferentes comportamentos em relação à sua não estacionariedade. Médio estresse atinge um valor máximo de INS mais elevado que Alto estresse. O Grito, por sua vez, atinge valores de INS bem mais elevados que as demais condições de estresse.

2.3 – Análise de Variações Acústicas Não Estacionárias

Para analisar variações acústicas não estacionárias, é necessário que se busquem os métodos mais adequados para esta tarefa, de modo que haja de forma adequada a detecção dessas variações. Isto pode auxiliar na definição de um atributo que seja interpretado como uma

impressão acústica dos estados afetivos. Entre as formas mais usuais de se analisar variações acústicas está a observação do sinal no domínio do tempo ou da frequência. A análise de Fourier é comumente utilizada quando se deseja observar propriedades espectrais de um sinal. Uma limitação da análise de Fourier é o fato de ela não ser adequada para sinais não estacionários [16]. A transformada Wavelet, diferentemente da transformada de Fourier, realiza a análise espectral de um sinal no domínio do tempo. Uma limitação da decomposição tempo-frequência Wavelet é o fato de ser utilizado bancos de filtros pré-definidos, o que pode levar a uma longa busca pelo banco de filtros ideal para cada tipo de sinal [52]. A EMD surgiu recentemente como uma técnica de análise tempo-frequência que, além de ser adequado para sinais não estacionários, possui as seguintes características:

- **Adaptatividade** – a decomposição depende exclusivamente do sinal (*data-driven*);
- **Localidade** – as funções intrínsecas de modo (IMFs – *Intrinsic Mode Functions*) são completamente baseadas nas propriedades locais do sinal;
- **Soma Perfeita (*Completeness*)** – a soma de todos os modos obtidos da EMD com o resíduo final garante a perfeita reconstrução do sinal original.

2.3.1 – Decomposição Empírica de Modos

A ideia principal da decomposição empírica de modos é analisar um sinal $x(t)$ contendo dois extremos consecutivos (máximo ou mínimo) nos pontos t_- e t_+, definindo uma componente de altas frequências do sinal, também chamada de detalhe $d(t)$, e uma componente de tendência local, ou resíduo $r(t)$, tal que

$$x(t) = d(t) + r(t)\,, \quad t_- \leq t \leq t_+\,. \tag{2.14}$$

O conjunto dos detalhes locais, obtidos de todos os extremos consecutivos de $x(t)$, compõe a primeira IMF (primeiro modo da decomposição). A separação entre componentes de altas e de baixas frequências é repetida de forma iterativa sobre o sinal residual $r(t)$, chegando-se a um conjunto de IMFs e a um resíduo de baixas frequências.

Para a realização da decomposição empírica de modos sobre um sinal $x(t)$, os seguintes passos são seguidos, de acordo com a Figura 2.6 [16, 53]. No início do algoritmo, para a primeira IMF ($k = 1$), $a_0(t) = x(t)$. Em seguida, são identificados todos os extremos de $a_{k-1}(t)$, ou seja, os pontos de máximo $x_{max}(t)$ e mínimo $x_{min}(t)$ locais. Após isso, obter as envoltórias $e_{max}(t)$ e $e_{min}(t)$, interpolando-se os pontos de máximo e de mínimo, respectivamente. Para isto, adota-se a interpolação polinomial de terceiro grau utilizando o método de *splines*. Em seguida, calcular a média entre as envoltórias: $a_k(t) = \big(e_{min}(t) + e_{max}(t)\big)\,/2$. Posteriormente, extrair as componentes de detalhes: $d(t) = x(t) - r(t)$. Para que o próximo passo seja realizado, a componente de detalhe $d(t)$ obtida deve obedecer a

duas propriedades: a primeira está relacionada à diferença entre a quantidade de extremos e a quantidade de cruzamento em zero, a qual deve ser nula ou igual a um; a outra propriedade diz respeito ao valor médio definido pelas envoltórias dos seus máximos e mínimos, o qual deve ser nulo. Caso contrário a estas propriedades, os passos até então são novamente efetuados, com $d(t)$ no lugar de $x(t)$. Este processo, denominado *sifting*, é repetido até garantir que a nova função $d(t)$ seja considerada uma IMF [16]. O próximo passo, então, é repetir a iteração sobre o sinal residual $r(t) = a_k(t)$.

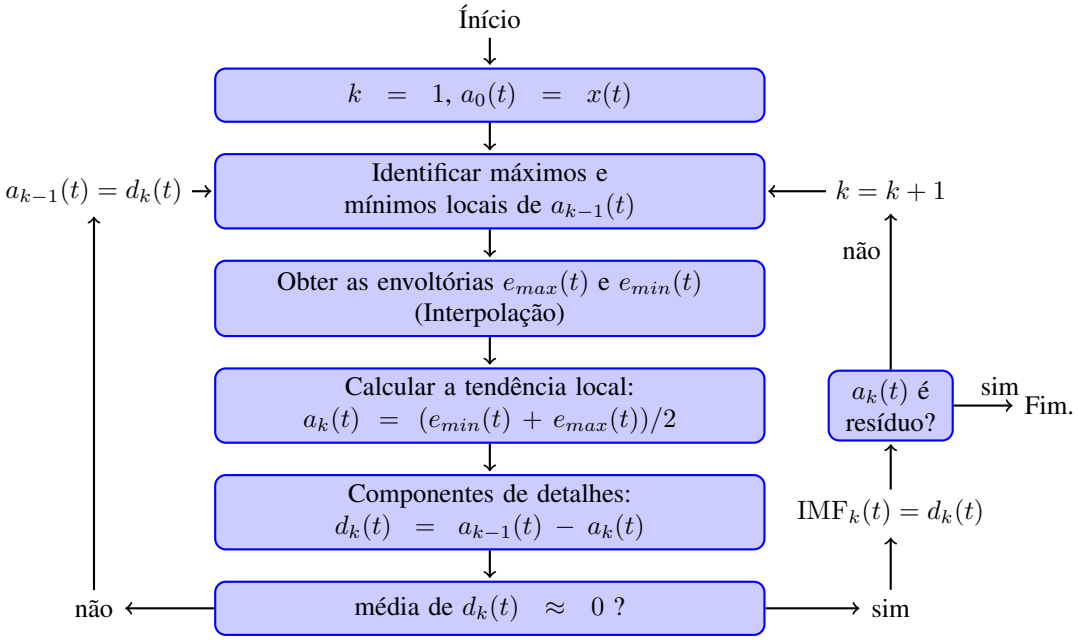


Figura 2.6 – Algoritmo do método EMD.

O algoritmo do método EMD assegura que qualquer sinal $x(t)$ pode ser decomposto em um número finito (K) de iterações, e pode ser escrito como

$$x(t) = \sum_{k=1}^{K} \text{IMF}_k(t) + r(t)\,, \tag{2.15}$$

em que $\text{IMF}_k(t), 1 \leq k \leq K$, são as funções de detalhes $d(t)$ obtidas no passo (4) de cada iteração, e $r(t)$ é o sinal residual final decorrente da última iteração.

Como mostrado em [53], a quantidade de cruzamentos por zero observada na decomposição empírica de modos pode ser uma indicação da frequência média de cada IMF, e a forma com que essa quantidade varia de um modo para outro traz evidências da estrutura hierárquica de um banco de filtros diádicos.

Exemplos de decomposição de 96 ms de três sinais de voz nos estados Neutro, Raiva e Tristeza são apresentados na Figura 2.7. A redução no número de extremos de um modo para o próximo implica que, localmente, as primeiras IMFs possuem oscilações mais rápidas (altas frequências) que as IMFs de maior índice. Isto significa que a EMD aplica uma separação alta-

frequência *versus* baixa-frequência entre as IMFs. A primeira IMF referente à Raiva possui amplitudes superiores àquelas dos outros estados emocionais, enquanto que na terceira IMF este papel passa a ser da emoção Tristeza. Pode ser notado, ainda, que nas mais baixas frequências (após a sexta IMF, por exemplo) a forma de onda já não contém tanta informação agregada ao sinal quanto as IMFs anteriores. Este é um indício de que a decomposição baseada em EMD enfatiza o conteúdo emocional presente na voz. No caso da Raiva (emoção de alta ativação), as variações acústicas não estacionárias estão mais concentradas nas IMFs de maior frequência. Por outro lado, Tristeza (emoção de baixa ativação) apresenta destaque nas IMFs de maior índice.

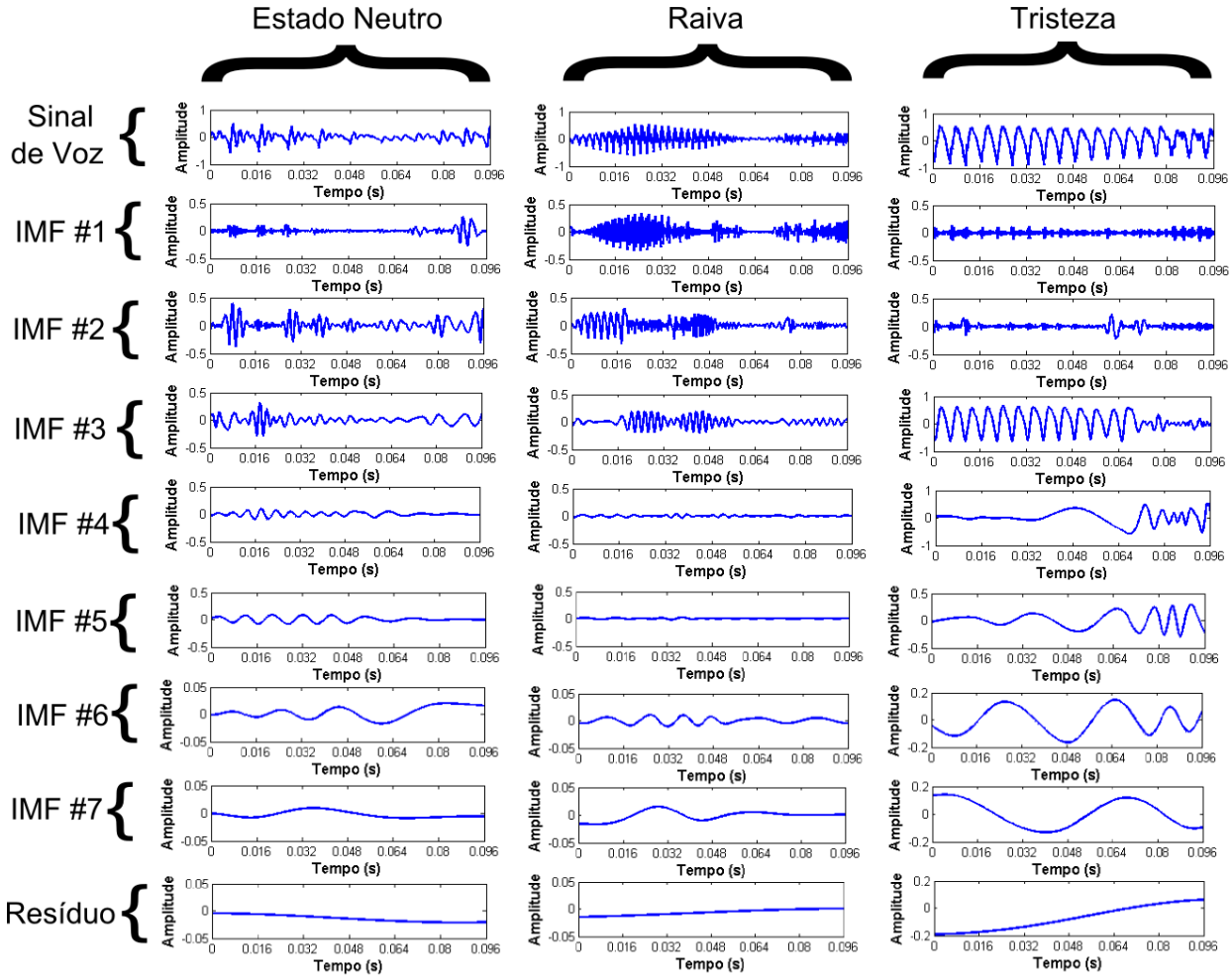


Figura 2.7 – EMD empregada em trechos de 96 ms de sinais de voz nos estados emocionais Neutro, Raiva e Tristeza.

Para uma análise mais objetiva as amplitudes das oscilações nas diferentes IMFs, foi calculada a Energia segmental em trechos de 96 ms de cinco emoções da base EMO-DB [54]: Raiva, Felicidade, Neutro, Tédio e Tristeza. Os valores médios da Energia por IMF são apresentados na Figura 2.8. Como pode ser verificado, nas frequências mais elevadas (primeira IMF), a emoção Raiva (alta ativação) apresenta a maior concentração de energia quando comparada à demais emoções. Por outro lado, a Tristeza (baixa ativação) tem este

comportamento nas frequências mais baixas (sexta IMF). Além disso, é possível notar que a partir da terceira IMF fica mais evidente a separabilidade entre as emoções de alta e baixa ativação. Isto representa uma relação com a densidade espectral de potência da emoções. Enquanto emoções de alta ativação apresentam os maiores valores de energia nas IMFs de menor índice, comportamento contrário ocorre com as emoções de baixa ativação.

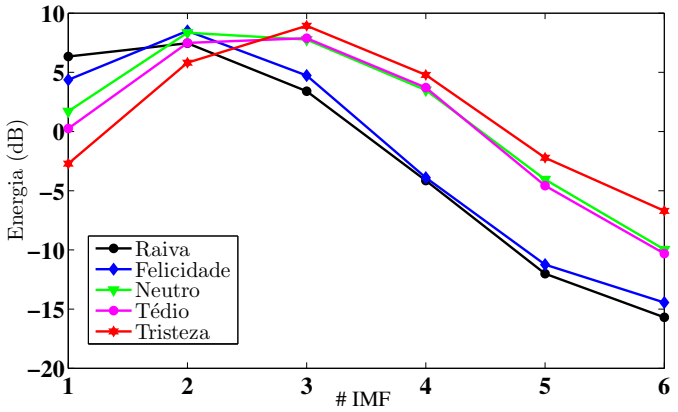


Figura 2.8 – Média dos valores de Energia (dB), por emoção, para cada uma das seis IMFs observadas.

A EMD apresenta algumas limitações que fizeram com que surgissem algumas abordagens que trazem soluções destes desafios em forma de variações do EMD [52, 55]. As principais questões relacionadas ao método EMD são as seguintes:

- **Problema de mistura entre modos (*mode mixing*)** – fenômeno que acontece quando uma IMF pode conter sinais de diferentes escalas ou quando sinais de escalas similares estão contidos em diferentes IMFs [56];
- **Cálculo da envoltória (interpolação)** – diz respeito a soluções de otimização para o procedimento de interpolação dos extremos no processo de decomposição [57].

Nesta Tese, a abordagem alternativa ao método EMD que foi escolhida é a EEMD (*Ensemble Empirical Mode Decomposition*), que procura solucionar o problema de mistura entre modos. Este estudo tem como objetivo verificar qual o método mais adequado para a tarefa de detecção das variações acústicas afetivas. Tal método é descrito a seguir.

• EEMD

Para evitar mistura entre modos (*mode mixing*), foi proposto o EEMD [56]. Este método utiliza ruído Gaussiano branco na decomposição EMD [53, 58].

A funções intrínsecas de modo no método EEMD, denotadas por $\overline{IMF}$, são definidas como sendo a média das IMFs correspondentes obtidas no EMD aplicado a uma quantidade I de sinais corrompidos $x^i(t)$, pela adição de sequências de ruído Gaussiano branco, $w^i(t)$, ao sinal original, $x(t)$. Assim,

$$x^i(t) = x(t) + w^i(t), \qquad i = 1, ..., I. \tag{2.16}$$

Os K modos $\text{IMF}_k^i(t)$ do sinal corrompido são obtidos da aplicação do EMD em $x^i(t)$. O k-ésimo modo de $x(t)$ é dado por:

$$\overline{IMF}_k = \frac{1}{I}\sum_{i=1}^{I} \text{IMF}_k^i(t). \tag{2.17}$$

Em termos gerais, a ideia básica do método EEMD é adicionar o ruído branco de forma que ele seja distribuído uniformemente em todo o espaço tempo-frequência. Com a estrutura similar a um banco de filtros[1] resultante da EMD, o ruído branco adicionado ao sinal posiciona as componentes de frequência de forma mais apropriada possível em uma escala mais adequada do que a EMD.

A utilização de um conjunto de sequências de ruído branco tem influência no refinamento da decomposição EEMD. Como mostrado em [56], quanto maior o valor de I, maior a possibilidade de as componentes de frequência serem detectadas na escala mais adequada. Como efeito da EEMD, as séries de ruído adicionadas se cancelam no resultado da Equação 2.17. Outro aspecto que pode ser observado na EEMD é a energia do ruído branco, em termos de sua variância (ou desvio padrão). Isto pode influenciar na detecção dos extremos durante a decomposição [56]. No desenvolvimento desta Tese, chegou-se a uma quantidade de I =100 realizações do ruído. Em termos de energia do ruído, diferentes valores de desvio padrão foram experimentados, entre 0,005 e 0,1.

Como exemplo de diferenças entre EMD e EEMD, na Figura 2.9 são apresentadas as referidas decomposições para um sinal de voz no estado Neutro. Note que nas diferentes faixas de frequência (IMFs) diferentes tipos de oscilações ocorrem para uma mesma IMF no caso da EMD (Figura 2.9a). Por outro lado, no caso da EEMD (Figura 2.9b), as oscilações ficam em escalas mais "apropriadas". Na segunda IMF, por exemplo, oscilações de mais baixa frequência (que poderiam estar em IMFs de maior índice) ocorrem em maior quantidade no caso da EMD, em comparação à EEMD. Para a quarta IMF, pouco depois de 100 ms na EMD as oscilações diferem do que seria mais adequado para aquela escala de frequência, o que é corrigido na quarta IMF com o método EEMD.

Outros trabalhos apresentam diferentes soluções ao problema de mistura entre modos, tais como a proposta do CEEMDAN (*Complete EEMD with Adaptative Noise*) [59] e do MEMD (*Multivariate EMD*) [60].

[1] A decomposição EMD pode ser considerada um banco de filtros diádico para o caso de ruído Gaussiano branco [56].

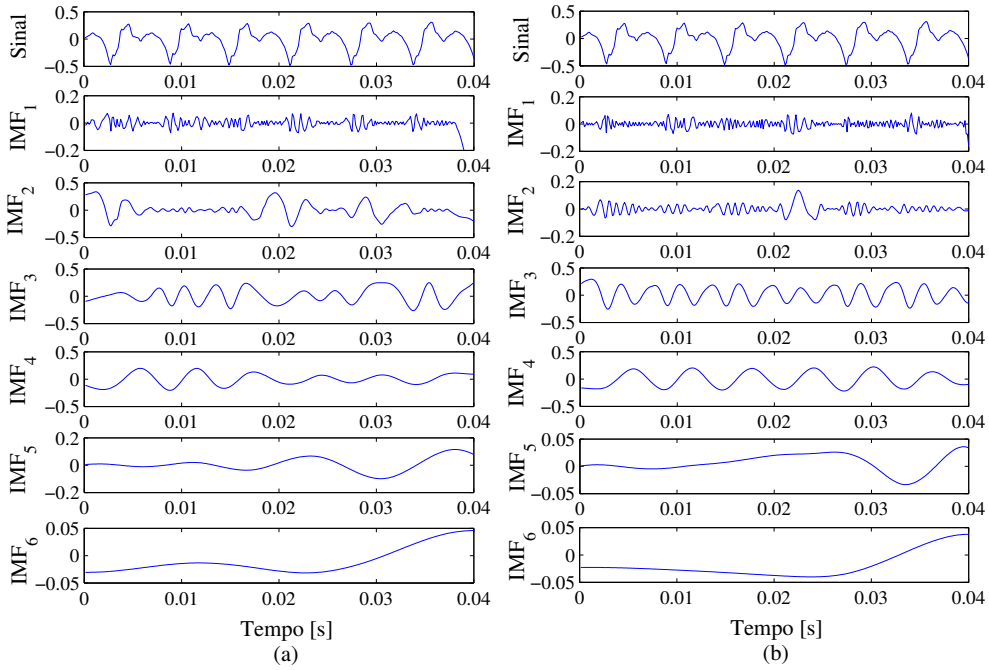


Figura 2.9 – Decomposição de um sinal de voz sob nenhuma emoção (estado Neutro) por meio de: (a) EMD e (b) EEMD.

2.4 – Resumo

Neste Capítulo foram apresentados conceitos relacionados a estados afetivos, com mais detalhe em emoções, que tem sido o principal objeto de estudo nesse contexto. Estudos como o de Darwin foram brevemente descritos, os quais serviram de base para as definições mais comuns de emoções: a abordagem discreta (que categoriza as emoções básicas) e a abordagem contínua (que agrupa as emoções em diferentes eixos). Os estados afetivos ainda foram apresentados como sendo variações acústicas não estacionárias. Para a análise dessas variações acústicas, dois métodos foram apresentados: o INS e a EMD. O uso do INS tem relevância neste trabalho no sentido de que as variações acústicas afetivas apresentam diferentes graus de não estacionariedade. Desta forma, este é um importante aspecto a ser levado em consideração. Como o método de decomposição baseado em EMD é adequado para sinais não estacionários, ele pode ser empregado na detecção dessas variações acústicas. INS e EMD (EEMD) foram as principais técnicas utilizadas nesta Tese para a análise das variações afetivas e para a definição de um novo atributo acústico, que é apresentado no Capítulo seguinte.

CAPÍTULO 3

HHHC: Atributo Acústico

Neste Capítulo, é proposto o vetor HHHC (*Hilbert-Huang-Hurst Coefficients*) como um novo atributo acústico não linear para a classificação de múltiplas emoções e condições de estresse. Também é proposto o método de extração do HHHC. Este novo atributo caracteriza as variações acústicas afetivas de acordo com seus efeitos na fonte de excitação vocal. Tais variações, analisadas no contexto de sua não estacionariedade, são destacadas por meio de um método adaptativo de decomposição tempo-frequência baseado na transformada de Hilbert-Huang (a EMD). O expoente de Hurst, que tem relação com a densidade espectral de potência das emoções, é estimado de cada componente obtida do processo de decomposição, formando assim o vetor HHHC. A motivação do HHHC é baseada nas seguintes razões: o significado físico do expoente de Hurst em capturar informação não linear dos estados afetivos a partir de seus efeitos na fonte de excitação; a detecção acústica dos estados afetivos realizada por meio da decomposição baseada em EMD; o fato de que estados afetivos levam diferentes graus de não estacionariedade ao sinal de voz; por se tratar de um atributo de fonte de excitação, é menos dependente do conteúdo linguístico do que outros tipos de atributos, a exemplo de atributos do trato vocal.

Para agregar informação ao atributo HHHC, é utilizado o INS. Como foi apresentado no Capítulo anterior, os estados afetivos podem ser considerados fontes acústicas que apresentam diferentes graus de não estacionariedade. Neste trabalho de Tese, o INS é proposto como medida da dinâmica não estacionária do sinal ao longo do tempo. Assim, a fusão HHHC+INS constitui-se de mais uma alternativa proposta para a tarefa de reconhecimento de emoções e condições de estresse.

Nas Seções seguintes, são descritos brevemente alguns dos atributos acústicos mais utilizados na literatura, os quais podem também ser definidos como sendo descritores de baixo nível. Em seguida, é apresentada uma análise dos métodos EMD e EEMD para a definição de qual o método mais ideal para enfatizar as variações acústicas afetivas. Então, é apresentado o expoente de Hurst e o método de extração do vetor HHHC.

3.1 – Atributos Acústicos

Um atributo acústico pode ser definido como sendo uma medida que caracteriza um fenômeno físico por meio da onda sonora resultante de um determinado processo (a produção da voz, por exemplo). A tarefa de análise e classificação de variações afetivas é uma das linhas de pesquisa mais recentes no contexto de processamento de voz, em comparação a tarefas como reconhecimento de fala e locutor. Por isso, atributos bem sucedidos nessas aplicações foram utilizados como ponto de partida nas pesquisas envolvendo reconhecimento de emoções [4]. Neste contexto, o desafio tem sido correlacionar medidas que descrevem o mecanismo de produção e modelagem do sinal de voz com os efeitos provocados por variações acústicas afetivas [9, 31].

A utilização dos atributos acústicos tem como ponto de partida o conjunto de aspectos que vão ser observados no sistema de produção vocal sob estados emocionais ou condições de estresse [4, 5]. Entre os atributos que podem ser encontrados na literatura, estão aqueles relacionados à fonte de excitação glotal [61], atributos relacionados ao trato vocal [6] e atributos relacionados ao modelo não-linear de produção da fala [12]. Ainda, há coleções de atributos que podem utilizar diferentes tipos de medidas para formar a matriz de atributos que é aplicada no processo de classificação [9].

3.1.1 – Atributos da Fonte de Excitação

Medidas que descrevem o comportamento da fonte de excitação glotal são comumente chamadas de atributos da fonte de excitação ou atributos da fonte vocal. Na análise de emoções, há medidas clássicas como o *Pitch* (ou frequência fundamental – F0) e atributos recentemente propostos, como é o caso do vetor pH.

• *Pitch*

O *Pitch*, ou frequência fundamental (F0), descreve a taxa de vibração das pregas vocais durante a fonação [61]. Para a estimação do *Pitch* a partir do sinal de voz, existem diferentes algoritmos, a exemplo do método da função de autocorrelação [62] e o método da função da média de diferenças de amplitudes (*Average Magnitude Difference Function* – AMDF) [63]. Geralmente, o *Pitch* é empregado juntamente com um conjunto de estatísticas obtidas dele, tais como média, valor máximo e valor mínimo [4, 29].

Uma vez que a tensão muscular da laringe pode ser afetada por variações emocionais, o *Pitch* é empregado a partir desta premissa. Porém, esta medida é mais promissora na separação de grupos de emoções, como por exemplo o caso emoções de alta ativação *versus* emoções de baixa ativação. Isto porque a F0 se comporta de maneira semelhante em diferentes emoções (como por exemplo Tédio e Tristeza) [4].

• Vetor pH

O vetor de coeficientes de Hurst (pH) foi proposto inicialmente em [64] como atributo para reconhecimento de locutor. Posteriormente, em [7] o vetor pH foi proposto como atributo para classificação de emoções e condições de estresse. O vetor pH está relacionado com as informações de excitação glotal.

Para a estimação do vetor de atributos pH, é empregado o extrator multi-dimensional baseado em *Wavelets* (M-dim-wav – *multi-dimensional wavelet-based estimator*), seguindo os seguintes passos:

- Decomposição Wavelet: aplica-se a transformada *Wavelet* discreta (DWT – *Discrete Wavelet Transform*) sucessivamente para decompor o sinal de voz em componentes de detalhe ($d(j,n)$) e aproximação ($a(j,n)$), em que j representa as escalas da decomposição ($j = 1, ..., J$) e n é o índice de cada escala. Os filtros propostos por [65] são utilizados na DWT;

- Estimação do expoente de Hurst (EH) [66]: para cada escala j, a variância dos coeficientes de detalhes é calculada por $\sigma_j^2 = (1/N_j) \sum_n d(j,n)^2$, em que N_j representa a quantidade de coeficientes da escala j. O expoente de Hurst do sinal de voz é estimado por $H_0 = (1+\theta)/2$, em que θ é a inclinação da reta obtida por regressão linear de $\log_2(\sigma_j^2)$ *versus* j.

- Extração do vetor pH: o vetor pH é composto por $J + 1$ valores de H ($H_0, H_1, ..., H_J$). O primeiro coeficiente, H_0, é obtido diretamente pela decomposição Wavelet do sinal de voz, conforme descrito no item acima. Os outros valores ($H_1, ..., H_J$) são obtidos aplicando-se novamente a decomposição Wavelet a cada uma das J sequências de detalhes e estimando novamente os valores de H.

3.1.2 – Atributos do Trato Vocal

Atributos do trato vocal capturam características do sinal de voz no domínio da frequência. Por isso, são também conhecidos como atributos espectrais [4, 67]. Em geral, este tipo de medida é extraído do sinal de voz em quadros de curta duração (20 ms a 32 ms), em que o sinal é considerado estacionário [68]. A compreensão para o uso deste tipo de atributo tem como ponto de partida o fato de que o trato vocal pode ter sua forma alterada sob o efeito de emoções [61], resultando, por exemplo, em impactos na distribuição de energia ao longo do espectro do sinal de voz [69]. Neste contexto, o atributo mais encontrado na literatura é baseado no vetor de coeficientes mel cepstrais (*Mel-Frequency Cepstrum Coefficients* – MFCC) [6, 7, 12]. Outros atributos espectrais foram recentemente propostos, tais como o vetor de coeficientes Gammatone (*Gammatone-Frequency Cepstral Coefficients* – GFCC) [70, 71], e vetor de parâmetros de Fourier (FP) [28]. Tais atributos são brevemente descritos nesta Seção.

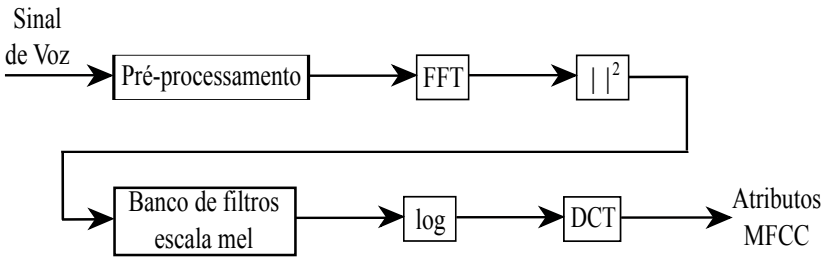


Figura 3.1 – Extração dos coeficientes MFCC.

Embora amplamente utilizados, os atributos espectrais são sensíveis às variações acústicas presentes nos sinais de voz [15].

• **MFCC**

Os atributos MFCC foram propostos e largamente utilizados em tarefas de reconhecimento de fala e de locutor [15, 27, 72, 73] e estão relacionados à percepção do ouvido humano. Tons puros ou sinais de voz não tem a percepção de suas frequências em uma escala linear, o que levou o desenvolvimento da chamada escala mel. Esta, por sua vez, aproxima computacionalmente a percepção auditiva [74, 75]. Como referência, 1 kHz equivale a 1.000 mels. A transformação de uma frequência f para a escala mel é descrita na Equação 3.1 [75].

$$Mel(f) = 1127\ln\left(1 + \frac{f}{700}\right). \tag{3.1}$$

Na análise mel cepstral, são utilizados bancos de filtros que tem por objetivo simular a resposta em frequência da membrana basilar do ouvido humano. No caso de sinais de voz, que são analisados até, aproximadamente, a frequência de 4 kHz, costumam ser utilizados 20 filtros triangulares com largura de banda de 300 mels, espaçados a 150 mels uns dos outros [74, 75].

A extração de atributos MFCC está ilustrada na Figura 3.1. A etapa de pré-processamento consiste na segmentação do sinal de voz em quadros de curta duração (20 ms – 30 ms). Em seguida, as amostras de cada quadro são levada ao domínio da frequência por meio da transformada rápida de Fourier (*Fast Fourier Trasform* – FFT), de onde é calculada a energia. O sinal transformado passa então por um banco de filtros na escala mel.

O conjunto de coeficientes MFCC (c_j) são obtidos de acordo com [72, 73]:

$$c_j = \sum_{i=1}^{F} (\log S_k) \cos\left[\frac{\pi j}{F}\left(k - \frac{1}{2}\right)\right], \qquad j = 1, 2, ..., D, \tag{3.2}$$

em que D é o número de coeficientes MFCC desejado, S_k é a potência na saída no k-ésimo filtro e F é o número de filtros na escala mel.

• GFCC

Os atributos GFCC foram propostos para tarefas de reconhecimento de locutor [76], e recentemente têm sido utilizados em reconhecimento de emoções [70, 71]. Assim como na análise cepstral que resulta nos coeficientes MFCC, a ideia geral para a obtenção dos coeficientes GFCC está baseada em uma aproximação computacional do sistema auditivo. Neste caso, são utilizados filtros *Gammatone*, os quais estão relacionados ao comportamento da coclea humana [77].

A resposta ao impulso de um filtro *Gammatone*, $g(t)$, é o produto da função distribuição Gamma e um sinal senoidal centrado na frequência f_c, de acordo com [78]

$$g(t) = Kt^{(n-1)}e^{-2\pi Bt}\cos(2\pi f_c + \varphi), \quad t > 0, \tag{3.3}$$

em que K é o fator de amplitude, n é a ordem do filtro, f_c é a frequência central em Hz, φ é a fase, e B representa a duração da resposta ao impulso. A extração dos atributos GFCC é semelhante àquela do MFCC até o passo da obtenção da FFT das amostras do sinal de voz. Após esta etapa, o sinal passa pelo banco de filtros *Gammatone* e então é empregada a transformada discreta do cosseno (*Discrete Cossine Transform* – DCT):

$$G_m = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} (\log Y_n) \cos\left[\frac{\pi m}{N}\left(n - \frac{1}{2}\right)\right], \quad 1 \leq m \leq M, \tag{3.4}$$

em que Y_n é a energia do sinal na n-ésima componente espectral, N é a quantidade de filtros *Gammatone*, e M é o número de atributos GFCC.

• FP

O vetor FP (*Fourier Parameters*) foi proposto em [28] para classificação de emoções baseado na transformada de Fourier (Equação 2.1) realizada quadro a quadro nos sinais de voz. As informações harmônicas $X[k]$ obtidas do sinal $x(t)$ representam as componentes espectrais modeladas no trato vocal. Para um conjunto de M harmônicas de $x(t)$, o vetor FP é dado por:

$$FP = [X_1[k], X_2[k], ..., X_p[k]], \quad 1 \leq p \leq M. \tag{3.5}$$

Embora seja estimado em quadros de curta duração de sinais de voz, o vetor FP foi tratado em [28] como um atributo global, ou seja, considerando apenas um valor da p-ésima harmônica para todo o sinal. Para tanto, após a extração do vetor FP dos segmentos do sinal de voz, é extraída a média desses valores. Outras estatísticas do vetor FP podem ser usadas, como mediana, máximo, mínimo e desvio padrão.

3.1.3 – Atributos Baseados no Operador TEO

O operador TEO (*Teager Energy Operation*) foi desenvolvido a partir dos estudos de Teager [32] e Kaiser [33] sobre fatores não lineares presentes no sistema de produção vocal. Eles mostraram que a voz é produzida pela interação não linear entre o fluxo de ar e vórtices formados dentro do trato vocal. Sob condições de emoção ou estresse, a tensão muscular no sistema de produção vocal afeta a interação fluxo-vórtice e, consequentemente, altera as características do sinal de voz [4]. Para um sinal de tempo discreto, $x[n]$, TEO ($\Psi\{\cdot\}$) é definido como:

$$\Psi\{x[n]\} = x^2[n] - x[n-1]x[n+1]. \tag{3.6}$$

Características de energia e frequência podem ser observadas com este operador. A frequência fundamental, por exemplo, muda em condições de estresse, assim como a distribuição de suas harmônicas [32, 79]. Além disso, sinais compostos por mais de uma frequência tem cada uma delas percebida pelo TEO, bem como suas interações entre si [12]. Zhou *et al.* [12] sugeriram medidas baseadas em TEO:

- Variação do componente de frequência modulada (**TEO-FM-Var**): uso baseado na observação de que variações suaves nos sinais de voz são devidas a efeitos de modulação;
- Área da envoltória da função de autocorrelação normalizada TEO (**TEO-Auto-Env**): partindo da observação que sinais de voz contêm componentes de amplitude e frequência modulada (AM e FM) centradas em frequências formantes, uma decomposição do sinal de voz em diferentes faixas de frequência pode detectar se a localização dos formantes é alterada em condições de estresse;
- Envelope de autocorrelação TEO baseado em banda crítica (**TEO-CB-Auto-Env**): o operador TEO é aplicado sobre diversas sub-bandas do sinal de voz para captar as variações na energia dos fluxos não-lineares para diferentes frequências de ressonância.

3.1.4 – Descritores de Baixo Nível e Funcionais

No universo de atributos acústicos utilizados em processamento de sinais de voz, podem ser encontradas duas categorias de medidas: os descritores de baixo nível (*Low-Level Descriptors* – LLDs) e os funcionais [9]. Os LLDs são atributos extraídos diretamente do sinal, em quadros de curta ou longa duração. Os funcionais, ou funções estatísticas, são medidos a partir dos LLDs, não tendo assim uma relação direta com o sinal [31]. Os atributos até então discutidos, como pH, MFCC e aqueles baseados no operador TEO são exemplos de LLDs. Os principais tipos de funcionais obtidos dos LLDs são a média, a mediana, os extremos (máximo e mínimo) e o desvio padrão [9].

Além de atributos da fonte de excitação, do trato vocal e daqueles baseados no operador TEO citados anteriormente, existem diversos tipos de LLDs utilizados em reconhecimento de emoções, dentre os quais podem ser citados os seguintes [4, 31, 80]:

- Energia – medida da intensidade sonora ao longo do tempo;
- Jitter – mede a variação de sucessivos períodos de *Pitch*;
- Shimmer – diferença das amplitudes de pico de períodos de *Pitch* consecutivos;
- HNR (*Harmonic-to-Noise Ratio*) – medida que indica a periodicidade de um sinal de voz por meio da relação entre as componentes periódicas (harmônicas) e as componentes não periódicas (ruído);
- Formantes – frequências de ressonância do filtro do trato vocal.

Em geral, quando diversos LLDs são agrupados após a extração, os funcionais são então as medidas empregadas no processo de classificação [9]. No caso de reconhecimento de estados afetivos, a ideia com isso tem sido agrupar atributos relacionados à produção da voz a fim de que seja aprimorada a identificação ou verificação do estado emocional ou condição de estresse [4]. Porém, o agrupamento de LLDs não significa, necessariamente, uma relação direta com a fisiologia dos estados afetivos. A fusão de informações de diferentes correlatos acústicos em conjuntos de características é um indício da dificuldade de se estabelecer um atributo acústico próprio para emoções e/ou condições de estresse.

3.1.5 – Discussão sobre os Atributos Acústicos

Diversos trabalhos relacionados a reconhecimento de estados afetivos podem ser encontrados na literatura. Todavia, diferentes pontos de vista são abordados nos estudos apresentados à comunidade científica. Dois principais pontos de partida se destacam neste contexto: trabalhos que tentam aprimorar a performance do classificador [29, 81, 82] e aqueles que priorizam a extração de LLDs, voltando-se à proposta de um atributo acústico de emoção [7, 28].

Em relação aos atributos acústicos dos sinais de voz, os LLDs mais comumente utilizados são os MFCC, principalmente para fins de comparação [7, 28, 83]. Embora amplamente utilizados devido ao sucesso em aplicações como reconhecimento de locutor, os MFCC apresentam performance inferior a outros atributos no caso de identificação de emoções e de condições de estresse [7, 25, 27, 28]. Por exemplo, em [7] o vetor pH (atributo da fonte de excitação) atingiu performance superior ao MFCC, evidenciando o fato de que atributos do trato vocal podem ser mais sensíveis ao conteúdo linguístico do que aqueles da fonte de excitação [15].

Alguns aspectos costumam ser levados em consideração para que seja definido um atributo acústico, entre os quais podem ser citados os seguintes [4]:

- Região de observação do sinal – relacionado à duração do segmento do sinal na extração de atributos e sua representação no processo de classificação. Em outras palavras, trata-se

da diferenciação entre medidas extraídas quadro a quadro (LLDs) e medidas globais (a exemplo dos funcionais) que representa a extração como sendo do sinal todo.

- Tipos de medidas – relacionado ao tipo de observação do sistema de produção vocal que vai ser levado em consideração na definição de um atributo. Neste contexto, por exemplo, há a diferenciação entre medidas da fonte de excitação e medidas do trato vocal.

- Coleção de atributos – relacionado ao conjunto de LLDs utilizados. Neste aspecto, pode ser observado a quantidade de coeficientes em um vetor de atributos [28] ou a fusão de diferentes atributos para fins de aumento de taxa de acerto [31].

Outro ponto importante a respeito das propostas apresentadas na literatura é que há trabalhos que têm focado na discriminação de emoções de forma agrupada [6, 31], ou seja, separando emoções dentro de dimensões como ativação e valência [42]. Neste sentido, um atributo ou um conjunto de atributos acústicos que sejam propostos podem não ser considerados, necessariamente, como atributos de emoção (ou de estresse), uma vez que não há neste caso uma representação dos estados afetivos de forma individual.

Apesar da existência de diversas abordagens sobre reconhecimento de emoções, que consideram todos os fatores e pontos de vista supracitados, ainda não há um consenso a respeito da definição de um atributo puro de emoções e/ou condições de estresse. Assim, o grande desafio é encontrar um medida acústica capaz de caracterizar os diferentes estados afetivos presentes no dia-a-dia do ser humano. A definição de uma "impressão afetiva vocal" tem, como consequência, a robustez nos sistemas de identificação de emoções e estresse. Um atributo acústico com esta característica pode ser eficiente em diferentes línguas e estilos de fala.

3.2 – Um Novo Atributo Acústico de Estados Afetivos

A principal contribuição desta Tese é a proposta de um novo atributo para a tarefa de classificação de emoções e condições de estresse. O vetor HHHC é um atributo não linear que caracteriza os efeitos das variações afetivas na fonte de excitação, sendo assim uma impressão acústica dessas variações. A decomposição baseada em EMD é utilizada na tarefa de detecção dos estados afetivos no sinal de voz. A partir da EMD, os coeficientes de Hurst (que são relacionados com a fonte de excitação) são estimados do processo de decomposição resultante para formar a matriz de atributos. As técnicas empregadas na composição do vetor HHHC são apresentadas a seguir, bem como é abordado o procedimento de extração deste atributo.

3.2.1 – EMD/EEMD

A decomposição baseada em EMD (como definido na Seção 2.3.1) é adaptativa, e é empregada nesta Tese com o objetivo de enfatizar as variações acústicas não estacionárias provocadas por emoções e condições de estresse. Além da EMD, é empregada a técnica

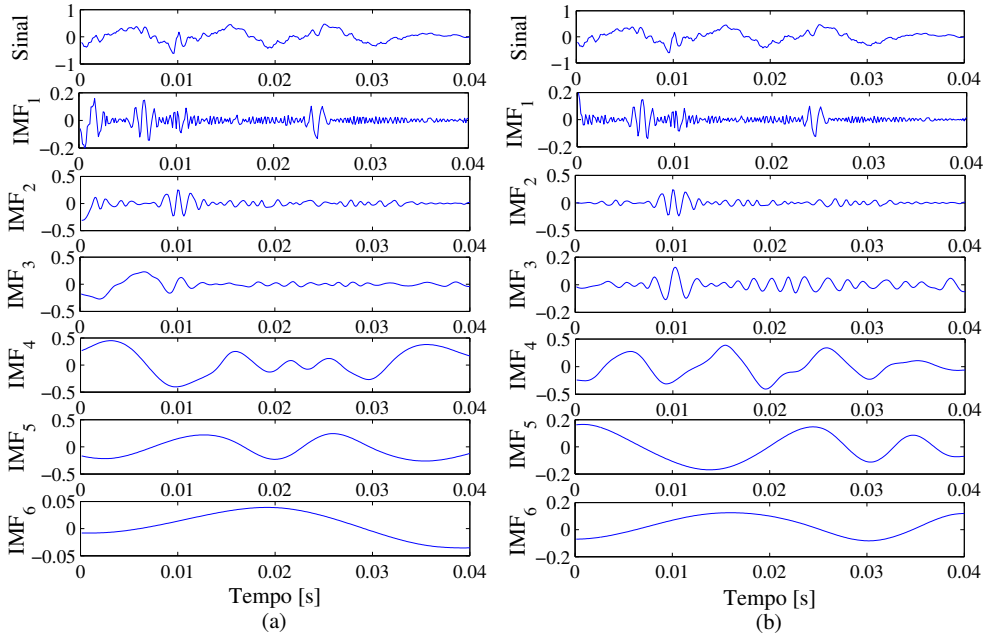


Figura 3.2 – Decomposição de um sinal de voz sob a emoção Tristeza por meio de: (a) EMD e (b) EEMD.

EEMD, como forma de investigar qual a abordagem que proporciona uma melhor detecção das variações acústicas afetivas. A partir de uma comparação entre EMD e EEMD é possível verificar, na prática, que o segundo método é mais eficaz em colocar, da forma mais fiel possível, as oscilações nas diferentes IMFs. Nos exemplo a seguir, a EEMD foi empregada utilizando ruído Gaussiano branco com desvio padrão igual a 0,01. Para o caso de análise em emoções, é empregada a base *Berlin Database of Emotional Speech* (EMO-DB) [54], e a base *Speech Under Simulated and Atual Stress* (SUSAS) [36] é utilizada na análise de condições de estresse.

• EMD versus EEMD em Diferentes Estados Emocionais

Na análise do estado de Tristeza, com decomposição apresentada na Figura 3.2, diferenças também podem ser observadas entre EMD e EEMD. Na primeira IMF, por exemplo, o método EEMD (Figura 3.2b) consegue corrigir o fenômeno do *mode mixing* nos primeiros 50 ms, em relação ao EEMD (Figura 3.2a). No caso da terceira IMF com a EMD, os primeiros 100 ms apresentam baixas oscilações e então, até 400 ms, é composto de frequências mais elevadas. Nesta mesma IMF com a EEMD, ocorrem oscilações de frequências mais próximas e mais adequadas à escala. Na quarta IMF, entre 100 ms e 300 ms, a EEMD apresenta oscilações com frequência mais baixa do que a mesma IMF no caso da EMD.

Na Figura 3.3 estão mostradas a EMD e a EEMD para um sinal sob o estado de Raiva. As diferenças mais perceptíveis entre os dois métodos podem ser notadas nas IMFs de maior índice. Com o método EEMD, por exemplo, a terceira IMF tem um problema de *mode mixing* corrigido entre os instantes 100 ms e 150 ms. Note que a partir da quarta IMF até a sexta a quantidade de picos e vales decaem de forma mais gradual e padronizada com a decomposição

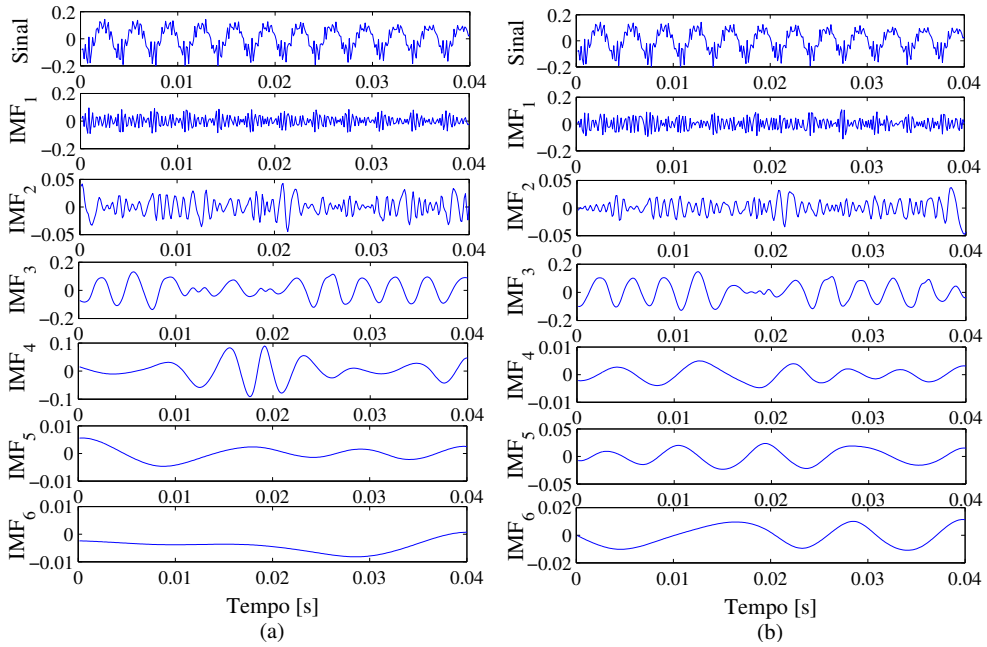


Figura 3.3 – Decomposição de um sinal de voz sob a emoção Raiva por meio de: (a) EMD e (b) EEMD.

EEMD do que com a EMD.

• EMD versus EEMD em Diferentes Condições de Estresse

Na análise com sinais da base SUSAS, são observadas as decomposições EMD e EEMD em um sinal no estado Neutro e outro na condição de alto estresse. Com outra base e em um contexto diferente daquele de estados emocionais, a análise com EEMD tende a ser mais promissora na detecção das variações acústicas não estacionárias.

Os métodos EMD e EEMD para um sinal no estado sem estresse da base SUSAS são apresentados na Figura 3.4. Assim como ocorreu no caso do estado Neutro para outra língua (Figura 2.9), as oscilações com a EEMD estão presentes em escalas mais adequadas. O *mode mixing* é corrigido nos primeiros 40 ms da primeira IMF. No caso da segunda IMF, oscilações mais apropriadas para esta escala são colocadas a partir de 150 ms com o método EEMD. Na quarta IMF com a EMD diferentes "padrões" de oscilação ocorrem antes e depois de 200 ms, o que não acontece com este índice de IMF no caso da EEMD.

Na Figura 3.5 são apresentadas as decomposições para um sinal acústico afetado por Alto estresse. A correção do *mode mixing* proporcionada pela EEMD pode ser mais claramente observada nos instantes 100 ms, 200 ms e 300 ms ao longo das IMFs. Oscilações de frequência mais baixa que aparecem nas primeiras três IMFs da EMD, ocorrem na quarta IMF do método EEMD.

De maneira geral, por meio das análise EMD *versus* EEMD entende-se que o método alternativo ao EMD pode destacar com mais precisão as variações acústicas introduzidas na voz por emoções e estresse. A partir de cada uma das IMFs obtidas é possível realizar algum

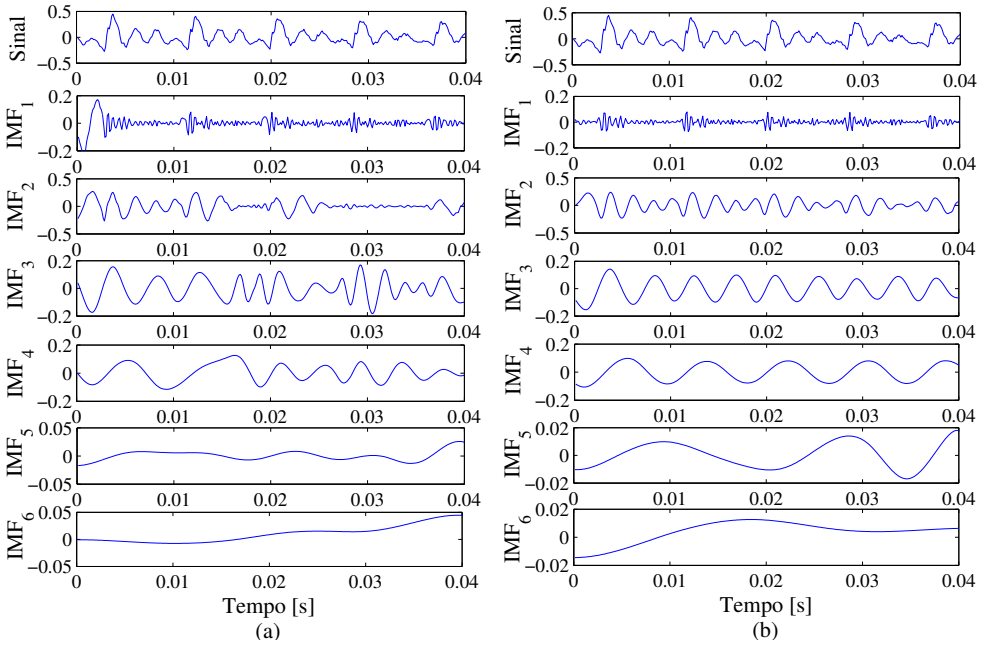


Figura 3.4 – Decomposição de um sinal de voz no estado Neutro (base SUSAS): (a) EMD e (b) EEMD.

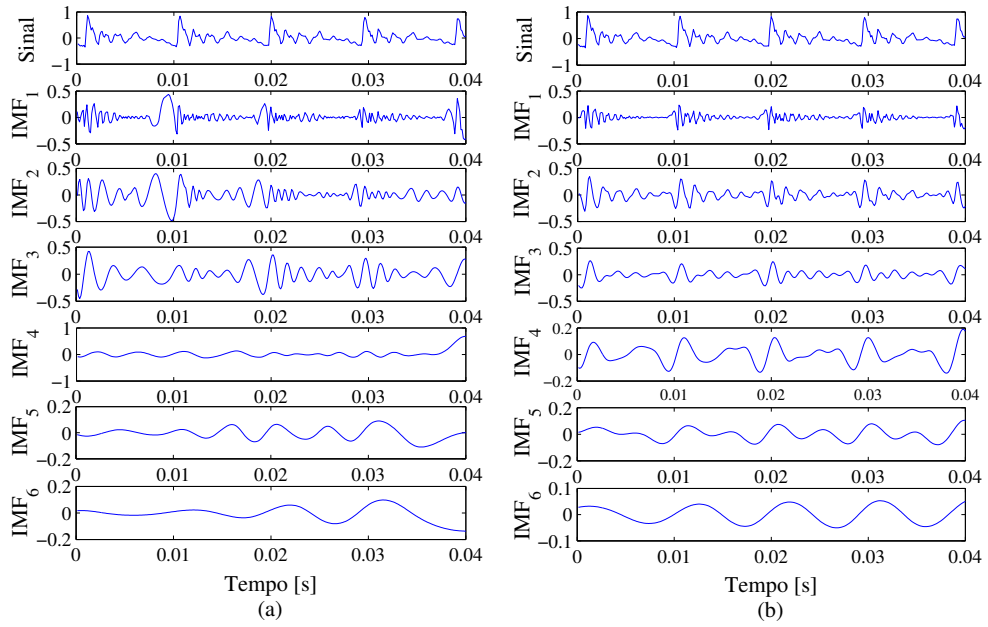


Figura 3.5 – Decomposição de um sinal de voz no estado de Alto estresse (base SUSAS): (a) EMD e (b) EEMD.

tipo de medida como forma de obter uma impressão acústica dessas variações. Neste trabalho de Tese, foi escolhido o expoente de Hurst, descrito a seguir, como conjunto de coeficientes para a formação de um novo atributo acústico, o HHHC.

3.2.2 – Coeficientes de Hurst

O expoente de Hurst ($0 \leq H \leq 1$), ou coeficiente de Hurst, expressa a dependência temporal ou grau de escala de um processo estocástico [17]. Esta medida também pode ser definida pela taxa de decaimento da função coeficiente de autocorrelação (*Autocorrelation Coefficient Function* – ACF) $\rho(k)$ ($-1 < \rho(k) < 1$) com $k \to \infty$. Seja um sinal de voz representado por um processo estocástico $x(t)$, com a função coeficiente de autocorrelação normalizada ($\rho(k)$) definida por

$$\rho(k) = \frac{E\left[\left(x(t) - \mu_x\right)\left(x(t+k) - \mu_x\right)\right]}{E\left[\left(x(t) - \mu_x\right)^2\right]}, \tag{3.7}$$

em que μ_x é a média de $x(t)$ e k é o atraso. O comportamento assintótico de $\rho(k)$ é dado por

$$\rho(k) \sim H(2H-1)k^{2(H-2)}, \quad k \to \infty. \tag{3.8}$$

Neste trabalho de Tese, os valores de H são estimados das IMFs de forma quadro a quadro utilizando um estimador baseado em wavelet (*wavelet-based estimator*) [66], o qual pode ser descrito em três passos principais:

1. Decomposição wavelet: a transformada wavelet discreta é aplicada para decompor sucessivamente a sequência de amostras de entrada em coeficientes[1] de aproximação ($a_w(j,n)$) e de detalhe ($d_w(j,n)$), em que j é a escala de decomposição ($j = 1, 2, ..., J$) e n é o índice de cada escala.

2. Estimação da Variância: para cada escala j, a variância $\sigma^2 = (1/N_j)\sum_n d_w(j,n)^2$ é calculada dos coeficientes de detalhe, em que N_j é o número de coeficientes disponíveis para cada escala j. Em [66], é mostrado que $E[\sigma_j^2] = C_H j^{2H-1}$, em que C_H é uma constante.

3. Cálculo do coeficiente de Hurst: uma regressão linear ponderada é usada para obter a inclinação θ da curva de $y_i = \log_2(\sigma_j^2)$ *versus* j. O coeficiente de Hurst é estimado como $H = (1+\theta)/2$.

O valor de H está relacionado com a densidade espectral de potência (*Power Spectral Density* – PSD) de um sinal $x(t)$. No contexto da produção vocal, as características espectrais do sinal acústico diferem de acordo com a excitação da fonte [14]. Em [7], foi mostrado que os valores de H estão relacionados com a PSD dos estados emocionais, como segue:

- Emoções de alta ativação (decaimento de -9 dB/oitava): ACF rapidamente decai a zero devido às componentes de alta frequência dominantes. Neste caso, $0 < H < 1/2$;

[1] O subescrito /w/ é usado para discriminar os componentes de detalhe ($d(t)$) e tendência ($a(t)$) do método EMD, dos coeficientes de detalhe ($d_w(j,n)$) e aproximação ($a_w(j,n)$) da decomposição wavelet.

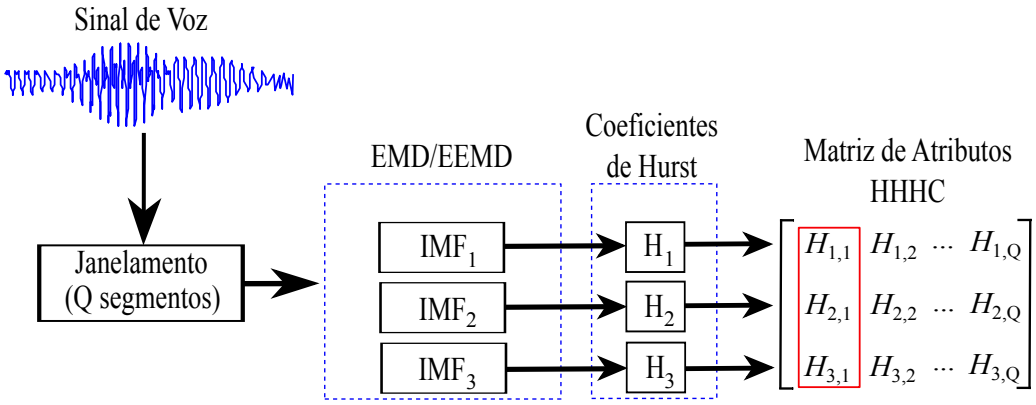


Figura 3.6 – Extração do vetor HHHC com três coeficientes.

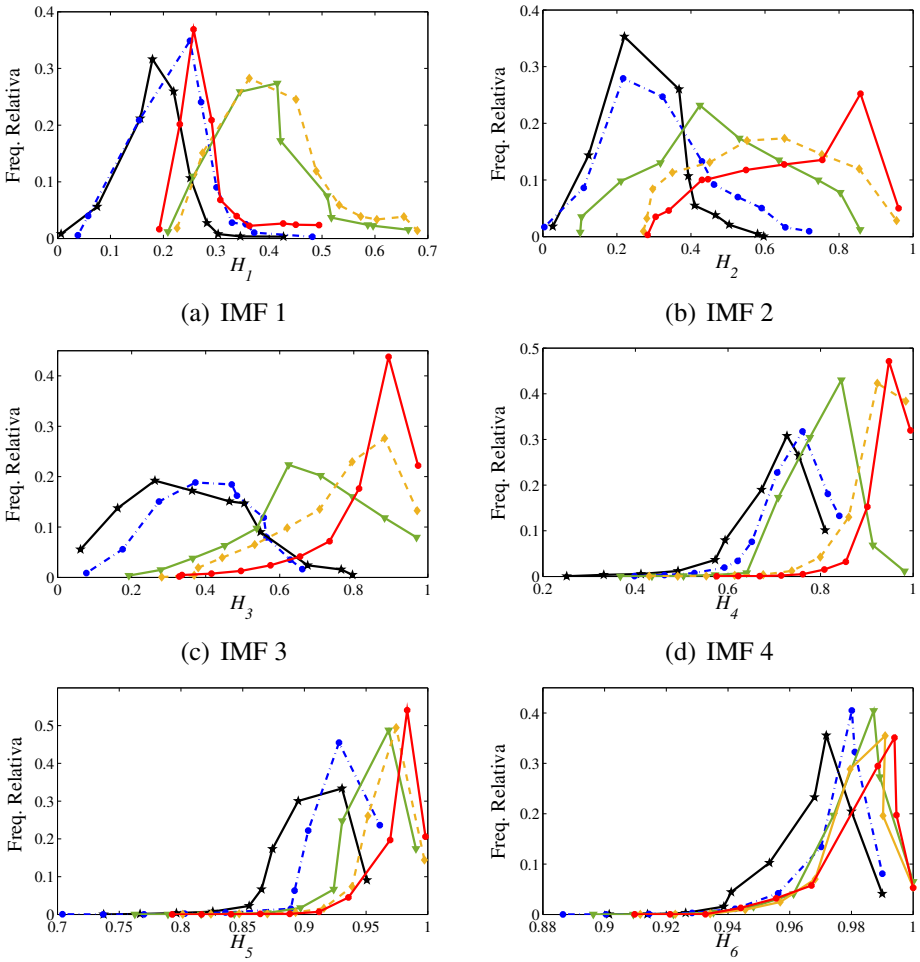
- Estado Neutro (decaimento de -12 dB/oitava): ACF geralmente exibe decaimento exponencial, em que $H \approx 1/2$;
- Emoções de baixa ativação (decaimento de -15 dB/oitava): ACF com decaimento mais lento devido às componentes de frequência com baixa energia. Assim, $1/2 < H < 1$.

De acordo com a relação supracitada entre o expoente de Hurst e emoções, ele foi escolhido neste trabalho de Tese para ser estimado a partir do resultado da decomposição baseada em EMD. Em [7], coeficientes de Hurst foram extraídos diretamente dos sinais acústicos, quadro a quadro, para a obtenção do vetor pH. Em contrapartida, nesta Tese, os coeficientes de Hurst são estimados a partir das IMFs extraídas dos sinais de voz.

3.2.3 – Método de Extração do HHHC

A extração do vetor HHHC de sinais acústicos com variações afetivas é realizado em duas etapas: decomposição pelo método EMD (ou EEMD); e uma estimação multicanal do expoente de Hurst. Um exemplo da estimação do vetor HHHC é mostrado na Figura 3.6. A decomposição baseada em EMD é aplicada no sinal de entrada, que pode ser um trecho sonoro do sinal de voz dividido em segmentos. Então, os valores de H são obtidos quadro a quadro de cada IMF. Neste caso, da IMF_1 a IMF_3, tem-se H_1 a H_3. Assim, o vetor HHHC $[H_1, H_2, H_3]$ é construído como um atributo acústico. Neste trabalho de Tese, são utilizados seis coeficientes de Hurst na formação da matriz de atributos do HHHC.

Na figura 3.7 são apresentados os histogramas obtidos da distribuição dos valores de H por cada IMF extraída de cinco estados emocionais da base EMO-DB: Raiva, Felicidade, Neutro, Tédio e Tristeza. Neste exemplo, a duração de cada sinal é de 40 s. Seis IMFs são obtidas por meio do método EEMD, aplicado em quadros de 80 ms com 50% de sobreposição. Os coeficientes de Hurst são calculados em segmentos de 20 ms (sem segmentação) dentro de cada IMF, usando wavelet com filtros Daubechies [65] com 12 coeficientes e escalas 3-12 no estimador baseado em wavelet. Note que, como as primeiras IMFs apresentam oscilações mais



(a) IMF 1 (b) IMF 2

(c) IMF 3 (d) IMF 4

(e) IMF 5 (f) IMF 6

Figura 3.7 – Distribuição dos valores dos coeficientes de Hurst para cada uma das seis IMF. Estados emocionais: Raiva (em preto), Felicidade (em azul), Neutro (em verde), Tédio (em amarelo) e Tristeza (em vermelho).

rápidas, os valores de H são mais baixos que aqueles obtidos das IMFs de maior índice. Nas últimas IMFs analisadas o valor de H é próximo da unidade. Apesar da separação entre altas e baixas frequências proporcionada pela EEMD, diferenças na frequência relativa dos valores de H podem ser observadas entre os estados emocionais. A separação mais evidente é entre emoções de alta ativação (Raiva e Felicidade) e emoções de baixa ativação (Tédio e Tristeza), principalmente a partir da segunda IMF (H_2 em diante). A ausência de emoção (estado Neutro) leva a valores de H que estão situados entre as emoções de alta e baixa ativação. Na segunda e na terceira IMF (Figuras 3.7b e 3.7c, respectivamente) é possível notar uma maior diferença, por exemplo, entre os picos dos histogramas obtidos para Raiva e Tristeza.

Os valores médios de H são apresentados na Figura 3.8. A medida em que se aumenta o índice da IMF, os valores de H tendem a unidade. A separação entre os estados emocionais pode, então, ser observada com mais evidência com a média de H. Assim, é possível notar que, uma vez que a EEMD enfatiza as variações acústicas, os coeficientes de Hurst capturam as características da fonte de excitação em cada IMF.

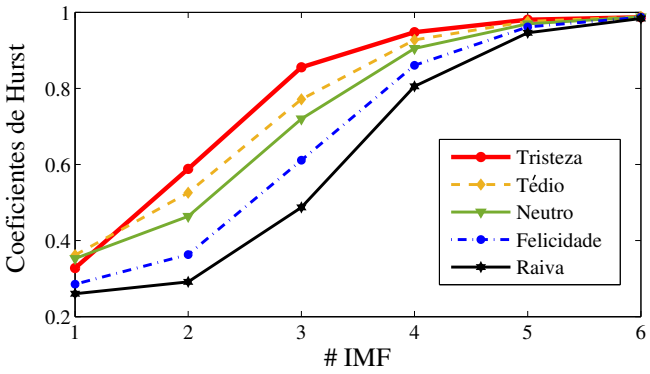
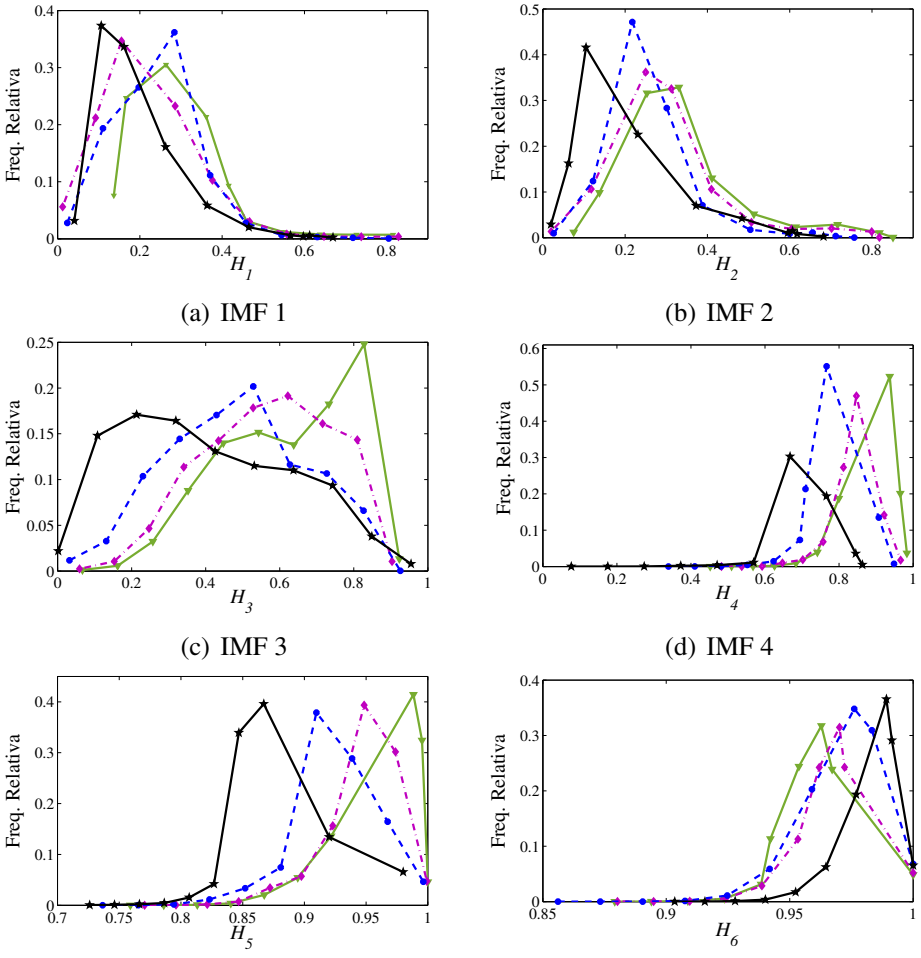


Figura 3.8 – Média dos coeficientes de Hurst de seis IMFs obtidas de sinais de voz com variações emocionais (base EMO-DB).



(a) IMF 1 (b) IMF 2

(c) IMF 3 (d) IMF 4

(e) IMF 5 (f) IMF 6

Figura 3.9 – Distribuição dos valores dos coeficientes de Hurst para cada uma das seis IMF. Condições de estresse: Grito (em preto), Alto estresse (em azul), Médio estresse (em violeta) e estado Neutro (em verde).

Histogramas obtidos dos valores de H para diferentes condições de estresse são apresentados na Figura 3.9. A extração do HHHC ocorreu da mesma forma para a obtenção da Figura 3.8. No caso da base SUSAS, foram utilizados 146 s de trechos sonoros dos sinais acústicos. Em todas as seis IMFs, é possível notar diferenças nas distribuições das condições de

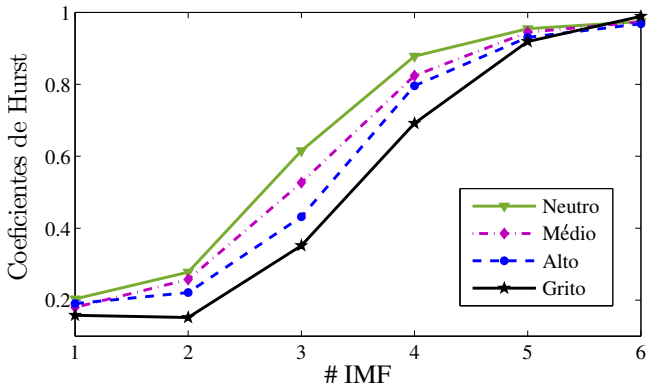


Figura 3.10 – Média dos coeficientes de Hurst de seis IMFs obtidas de sinais de voz com variações de condições de estresse (base SUSAS).

estresse. Além da diferença entre presença e ausência de estresse, note que os graus de estresse se diferem de acordo com o coeficiente H. Na terceira IMF ocorre um maior espalhamento na distribuição dos valores de H, bem como a separação entre os picos de frequência relativa dos estados Neutro e Grito é maior em relação às outras IMFs.

Na Figura 3.10 são apresentados os valores médios de H por IMF. As curvas para os estados de estresse aparecem, no gráfico, abaixo da curva para o estado Neutro. Isto indica a presença de componentes de alta frequência nos sinais sob estresse, resultando em menores valores de H em cada IMF em relação ao estado Neutro. Como foi observado com os histogramas da terceira IMF (Figura 3.9c), é possível notar a maior distância entre as médias de H das condições de estresse. Ainda, note que da primeira à quarta IMF, as médias para o estado de Grito é mais distante dos demais estados de estresse.

3.2.4 – Análise de Separabilidade do Atributo HHHC

Como foi observado na Seção 3.2.1, o método EEMD procura evitar o fenômeno do *mode mixing* e, dessa forma, tende a ser mais eficiente para destacar as variações acústicas não estacionárias provocadas por emoções e estresse. Para investigar o potencial de detecção da EMD e da EEMD, o vetor HHHC foi extraído utilizando ambas as abordagens. A EEMD foi realizada considerando um nível de ruído branco Gaussiano com desvio-padrão igual a 0,01. Então, utilizou-se a distância de Battacharyya como medida de separabilidade entre os estados afetivos, de forma a verificar qual a decomposição mais promissora na formação do vetor HHHC.

• Distância de Bhattacharyya

Dadas duas distribuições de probabilidade, $p_1(x)$ e $p_2(x)$, a distância de Bhattacharyya (Bd - *Bhattacharyya distance*) mede a similaridade entre elas [84]. Para o cálculo dessa distância, primeiramente é medido o chamado coeficiente de Bhattacharyya, dado por:

$$\rho = \int \sqrt{p_1(x)p_2(x)}\, dx, \tag{3.9}$$

em que $0 < \rho < 1$ quantifica quão sobrepostas estão as duas distribuições. Assim, a distância de Bhattacharyya é calculada como sendo:

$$Bd = -\ln\, \rho, \tag{3.10}$$

em que $0 < Bd < \infty$.

• *Bd* para HHHC na Análise de Emoções

Nas Tabelas 3.1 e 3.2 estão apresentados os valores de Bd entre cada um dos estados emocionais da base EMO-DB, considerando o vetor HHHC formado a partir das decomposições com EMD e EEMD, respectivamente. Os valores destacados mostram em qual IMF há a maior separação entre as variações emocionais. No caso da EMD, em todos os cenários a terceira IMF é mais discriminativa, exceto para o caso de separação entre Neutro e Tédio, que por sua vez apresenta maior Bd na quarta IMF. Note que na maioria das IMFs este cenário apresenta os menores valores de Bd, indicando a maior semelhança entre esses estados afetivos. Ainda, pode ser verificado a relação do HHHC com a PSD dos estados emocionais. As maiores Bd foram obtidas nos cenários com emoções com níveis de ativação distintos (ex: Raiva *versus* Tristeza).

No contexto da decomposição com EEMD (Tabela 3.2), a terceira IMF foi a mais discriminativa na maioria dos cenários, exceto no caso Neutro *versus* Felicidade. Em todos os cenários, os valores de Bd foram maiores do que aqueles obtidos utilizando a EMD. Além disso, a relação entre HHHC e PSD também pode ser observada, com menores valores de Bd para emoções com nível de ativação semelhante (ex: Raiva *versus* Felicidade) e maiores valores de Bd para emoções distintas de acordo com a ativação.

Tabela 3.1 – Distância de Battacharyya para os componentes do vetor HHHC baseado em EMD para a base EMO-DB.

Cenário	H_1	H_2	H_3	H_4	H_5	H_6
Raiva × Tristeza	0,046	0,287	**0,592**	0,409	0,182	0,011
Raiva × Tédio	0,075	0,219	**0,311**	0,259	0,131	0,015
Raiva × Felicidade	0,003	0,015	**0,038**	0,019	0,013	0,009
Felicidade × Tristeza	0,026	0,140	**0,267**	0,188	0,066	0,003
Felicidade × Tédio	0,037	0,088	**0,097**	0,091	0,038	0,002
Neutro × Raiva	0,068	0,127	**0,188**	0,091	0,032	0,013
Neutro × Felicidade	0,034	0,037	**0,043**	0,014	0,011	0,001
Neutro × Tédio	0,002	0,012	0,013	**0,015**	0,009	0,001
Neutro × Tristeza	0,020	0,048	**0,120**	0,071	0,027	0,002
Tristeza × Tédio	0,006	0,029	**0,054**	0,019	0,005	0,001

Tabela 3.2 – Distância de Battacharyya para os componentes do vetor HHHC baseado em EEMD para a base EMO-DB.

Cenário	H_1	H_2	H_3	H_4	H_5	H_6
Raiva × Tristeza	0,113	0,308	**0,754**	0,557	0,234	0,017
Raiva × Tédio	0,146	0,119	**0,417**	0,303	0,053	0,001
Raiva × Felicidade	0,006	0,022	**0,041**	0,023	0,018	0,001
Felicidade × Tristeza	0,104	0,187	**0,438**	0,207	0,052	0,021
Felicidade × Tédio	0,134	0,155	**0,189**	0,110	0,027	0,012
Neutro × Raiva	0,135	0,119	**0,157**	0,088	0,044	0,001
Neutro × Felicidade	**0,145**	0,049	0,039	0,021	0,003	0,003
Neutro × Tédio	0,015	0,056	**0,071**	0,037	0,019	0,010
Neutro × Tristeza	0,023	0,094	**0,258**	0,101	0,091	0,017
Tristeza × Tédio	0,008	0,035	**0,066**	0,014	0,051	0,001

• *Bd* para HHHC na Análise de Condições de Estresse

Nas Tabelas 3.3 e 3.4 estão apresentados os valores de Bd entre os cenários de condições de estresse (base SUSAS) com o vetor HHHC formado por meio da decomposição com EMD e EEMD, respectivamente. No contexto da utilização da EMD, na maioria dos cenários a terceira IMF mostrou-se como sendo a mais discriminante, exceto nos casos Médio estresse *versus* Grito e Grito *versus* Alto estresse. A maior separação pode ser observada no cenário Neutro *versus* Grito ($Bd = 0,152$), enquanto que os menores valores de Bd são observados na discriminação entre o estado Neutro e o Médio estresse. No caso da utilização da decomposição com EEMD, exceto o cenário Médio *versus* Alto (que obteve a maior separabilidade na primeira IMF), os maiores valores de Bd estão concentrados entre a terceira e quarta IMF. Também para a base SUSAS, os maiores valores de distância foram obtidos para os cenários em que o HHHC foi extraído por meio da utilização da EEMD.

Tabela 3.3 – Distância de Battacharyya para os componentes do vetor HHHC baseado em EMD para a base SUSAS.

Cenário	H_1	H_2	H_3	H_4	H_5	H_6
Neutro×Grito	0,036	0,008	**0,152**	0,006	0,008	0,006
Neutro×Alto	0,005	0,003	**0,103**	0,073	0,031	0,001
Neutro×Médio	0,001	0,006	**0,014**	0,012	0,003	0,001
Médio×Alto	0,001	0,013	**0,049**	0,034	0,016	0,003
Médio×Grito	0,015	**0,068**	0,027	0,028	0,009	0,001
Grito×Alto	0,045	0,077	0,02	**0,100**	0,006	0,002

3.2.5 – HHHC+INS

A proposta do vetor HHHC leva em consideração a não estacionariedade provocada pelas variações acústicas afetivas na fonte de excitação. Por meio da análise dessas variações, utilizando o INS, foi possível observar que as variações acústicas afetivas não estacionárias apresentam graus de não estacionariedade diferentes. Essas diferenças de não estacionariedade permitem o expoente de Husrt capture informações não lineares em cada uma das IMFs obtidas

Tabela 3.4 – Distância de Battacharyya para os componentes do vetor HHHC baseado em EEMD para a base SUSAS.

Cenário	H_1	H_2	H_3	H_4	H_5	H_6
Neutro×Grito	0,052	0,025	0,080	**0,213**	0,035	0,005
Neutro×Alto	0,040	0,092	**0,147**	0,137	0,023	0,022
Neutro×Médio	0,020	0,025	**0,080**	0,068	0,014	0,007
Médio×Alto	**0,057**	0,027	0,021	0,015	0,007	0,011
Médio×Grito	0,015	0,082	**0,140**	0,096	0,007	0,006
Grito×Alto	0,079	0,023	0,090	**0,135**	0,005	0,008

da decomposição baseada em EMD, que estão relacionadas com a PSD do estados emocionais.

Além da contribuição do INS na análise da não estacionariedade das variações acústicas afetivas, seu uso foi também investigado neste trabalho de Tese como informação adicional à matriz de atributos do HHHC, formando assim o HHHC+INS. Isto significa que, além do vetor HHHC para a classificação de estados afetivos, a fusão HHHC+INS também é proposta como informação complementar. Uma vez que o método EEMD demonstrou-se como sendo mais promissor na formação do vetor HHHC, ele foi utilizado na obtenção das IMFs obtidas dos sinais de voz das bases acústicas. Então, o INS foi estimado em cada IMF considerando 10 escalas de observação diferentes (T_h/T): 0,0015, 0,025, 0,05, 0,1, 0,15, 0,2, 0,25, 0,3, 0,4 e 0,5.

3.3 – Resumo

Neste Capítulo foi definido o vetor HHHC como novo atributo acústico, proposto para a classificação de variações afetivas não estacionárias. Inicialmente, foram apresentados os principais tipos de atributos acústicos empregados na tarefa de reconhecimento de emoções e condições de estresse. Foram apresentados atributos da fonte de excitação, do trato vocal e atributos baseados no operador TEO. Como foi discutido, não existe ainda um atributo acústico puro e estabelecido para o caso de classificação de emoções e estresse. Em seguida, na proposta do novo atributo acústico, foi apresentada uma análise da diferença entre os métodos EMD e EEMD na tarefa de detecção das variações acústicas não estacionárias. Neste sentido, foi observado que a EEMD é mais eficaz da tarefa de destaque dessas variações, podendo ser assim mais robusta para o atributo HHHC. O expoente de Hurst, que tem relação com a PSD dos estados emocionais, foi escolhido como coeficiente a ser estimado nas IMFs obtidas da decomposição baseada em EMD. Experimentos foram realizados com a distância de Battacharyya a fim de definir o método mais apropriado para a decomposição dos sinais acústicos: EMD ou EEMD, o que fortaleceu a ideia de que a EEMD é mais promissora na separação das variações acústicas afetivas. Adicionalmente, foi brevemente descrita a inclusão do INS como informação complementar na matriz de artibutos do HHHC, formando assim o HHHC+INS. No Capítulo seguinte, são apresentados os resultados obtidos no procedimento de

classificação das variações acústicas utilizando o vetor HHHC.

CAPÍTULO 4

Classificação das Variações Acústicas Afetivas Não Estacionárias

Neste Capítulo, são apresentados os experimentos concernentes à classificação das variações acústicas afetivas não estacionárias. Neste contexto, são empregados métodos clássicos de classificação de padrões relacionados a abordagens estocásticas e abordagens não paramétricas. Adicionalmente, é proposto o α-GMM como mais uma técnica de classificação no contexto de emoções e condições de estresse. Dessa forma, é avaliada a robustez do atributo acústico HHHC e sua fusão com a informação adicional do INS (HHHC+INS). Como forma de comparação, o vetor pH, os coeficientes MFCC e o atributo TEO são examinados nos experimentos. Ainda, é analisada a fusão do vetor HHHC com cada um desses atributos a fim de se verificar a sua contribuição na acurácia dos mesmos.

A seguir, são descritos brevemente os métodos de classificação utilizados nesta Tese, a metodologia empregada nos experimentos, os resultados da classificação considerando as cinco bases acústicas com variações afetivas e, por fim, um resumo do Capítulo.

4.1 – Métodos de Classificação

Ao longo do desenvolvimento desta Tese, buscou-se empregar os classificadores mais comuns na literatura no contexto de reconhecimento de estados afetivos. Assim, Foram utilizadas técnicas de classificação estocásticas e métodos baseados em aprendizado de máquina. No contexto do classificadores estocásticos clássicos, estão os Modelos de Misturas Gaussianas (*Gaussian Mixture Models* – GMM) [20] e os Modelos de Markov Escondidos (*Hidden Markov Models* – HMM) [21]. A técnica de Máquinas de Vetor de Suporte (*Support Vector Machines* – SVM) [22] foi empregada como método não paramétrico. Estas técnicas são encontradas em muitos trabalhos da literatura de reconhecimento de emoções e estresse [4, 7, 9, 31, 85].

Como contribuição desta Tese, é proposto o α-GMM, que assim como GMM e HMM,

consiste de uma abordagem estocástica. Para estes tipos de modelos, cada estado afetivo é modelado como sendo uma fonte probabilística, ou seja, a distribuição do atributo acústico é estimada dentro de cada estado emocional ou cada condição de estresse. Em geral, a classificação é realizada a partir do cálculo da verossimilhança entre o sinal de teste e os modelos [21]. Por outro lado, na abordagem com o SVM, o classificador modela a fronteira entre os estados afetivos, e a dissimilaridade é avaliada por meio de alguma medida de distorção [86].

4.1.1 – Métodos Estocásticos

• GMM

O modelo estocástico GMM foi proposto para reconhecimento de locutor [20] e é amplamente utilizado na literatura para classificação em diversas aplicações de sinais de voz [4]. Como definição, o GMM é uma soma ponderada de G componentes Gaussianas,

$$p\left(\mathbf{x} \mid \lambda\right) = \sum_{g=1}^{G} w_g b_g\left(\mathbf{x}\right), \tag{4.1}$$

em que:

- **x** é um vetor de atributos com D elementos;
- w_g $(g = 1, 2, ..., G)$ são os pesos das componentes Gaussianas;
- $b_g\left(\mathbf{x}\right)$ são componentes Gaussianas com vetor média $\vec{\mu}_g$ e matriz de covariância K_g, representadas da seguinte forma:

$$b_g\left(\mathbf{x}\right) = \frac{1}{\left(2\pi\right)^{\frac{D}{2}}\sqrt{\det K_g}} \exp\left(-\frac{1}{2}\left(\mathbf{x} - \vec{\mu_g}\right)^T K_g^{-1}\left(\mathbf{x} - \vec{\mu_g}\right)\right) \tag{4.2}$$

Para cada estado afetivo ε, é gerado um modelo GMM que é caracterizado pelos pesos, vetor média e matriz covariância:

$$\lambda_\varepsilon = \{w_g, \vec{\mu_g}, K_g\}, \qquad g = 1, 2, ..., G. \tag{4.3}$$

Na primeira fase da etapa de classificação, o treinamento, os modelos das variações acústicas afetivas não estacionárias são gerados a partir da matriz $\mathbf{X}_{Q\times D}$ de atributos, utilizando o algoritmo EM (*Expectation Maximization*) [20]. Dessa forma, o classificador busca um modelo λ que maximize a verosimilhança entre seus parâmetros e a matriz de atributos,

$$p\left(\mathbf{X} \mid \lambda_\varepsilon\right) = \prod_{t=1}^{Q} p\left(\mathbf{x}_t \mid \lambda_\varepsilon\right). \tag{4.4}$$

Na etapa de testes, dada uma matriz $\mathbf{X}_{Q\times D}$ extraída do sinal de voz de teste, o estado afetivo identificado é aquele cujo modelo λ maximiza a verosimilhança da Equação 4.4.

• HMM

O HMM, proposto inicialmente para reconhecimento de voz [21], consiste de um conjunto finito de estados internos que geram uma série de eventos externos (observações). Estes estados estão escondidos do observador. Os estados escondidos do modelo capturam a estrutura temporal de um sinal com variações acústicas. Matematicamente, o método HMM pode ser caracterizado por três problemas fundamentais:

1. Probabilidade: Dado um HMM $\lambda_L = (A, B)$ com K estados, e uma sequência de observações $\mathbf{x}$, determine a probabilidade $p(\mathbf{x}|\lambda)$, em que A é uma matriz de probabilidades de transição a_{jk}, $j, k = 1, 2, ..., K$, do estado j ao estado k, e B é o conjunto de densidades b_j;

2. Decodificação: Dado uma sequência de observações $\mathbf{x}$ e um HMM λ_L, descubra a sequência de estados escondidos;

3. Aprendizado: Dado uma sequência de observações $\mathbf{x}$ e um conjunto de estados no HMM, aprenda os parâmetros A e B.

O algoritmo padrão para o treinamento do HMM é o *forward-backward*, ou algoritmo de Baum-Welch [87]. Ele obtém as matrizes A e B que maximizam a probabilidade $p(\mathbf{x}|\lambda)$. O algoritmo de Viterbi é comumente utilizado para a decodificação [88].

4.1.2 – Método de Aprendizado de Máquina

• SVM

SVM [22] é um método de aprendizado de máquina supervisionado amplamente utilizado em classificação de dados. A ideia geral do SVM é encontrar o hiperplano de separação ótimo que maximize a margem nos dados de treinamento. Para este fim, esta técnica transforma vetores de entrada em um espaço de características de alta dimensão, usando uma transformação não linear (com uma função *kernel*), e então realiza uma separação linear neste espaço de características. A partir de um conjunto de treinamento u_ξ, $\{u_\xi\}_{\xi=1}^{N} = \{(\mathbf{x}_\xi, L_\xi)\}_{\xi=1}^{N}$, $L_\xi \in \{-1, +1\}$ representa o estado afetivo L do sinal acústico representado por ξ. Assim, o classificador é um hiperplano definido como

$$g(\mathbf{x}) = \boldsymbol{w}^{\mathrm{T}}\mathbf{x} + b, \tag{4.5}$$

em que $\boldsymbol{w}$ é o vetor normal que é perpendicular ao hiperplano, e b é o deslocamento do hiperplano a partir da origem. Dessa forma,

$$\begin{aligned} \boldsymbol{w}^{\mathrm{T}}\mathbf{x} + b \geq 0 \quad , \quad L_\xi = +1 \\ \boldsymbol{w}^{\mathrm{T}}\mathbf{x} + b < 0 \quad , \quad L_\xi = -1 \end{aligned} \tag{4.6}$$

Então, o hiperplano é escolhido por meio da solução do seguinte problema de otimização:

$$\text{minimize } \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}, \tag{4.7}$$

que está sujeito a

$$L_\xi \left(\boldsymbol{w}^{\mathrm{T}}\mathbf{x} + b\right) \geq 1, \quad \xi = 1, 2, ..., N. \tag{4.8}$$

Neste trabalho, os dados de entrada do classificador SVM são obtidos a partir dos vetores de média das matrizes de atributos. Esta estatística tem sido mais promissora que outras, tais como mediana e valor máximo, como foi observado em [89]. O *kernel* utilizado foi a Função de Bases Radiais (*Radial Basis Function* – RBF). Para a classificação multiestilo (com mais de duas classes), foi empregada a estratégia *one-versus-one*, a qual cria $k(k-1)/2$ classificadores binários (k classes) [90, 91].

4.1.3 – Proposta do α-GMM para a Classificação de Variações Acústicas Afetivas

Nesta Tese, além do uso dos classificadores estocásticos clássicos (GMM e HMM) e do SVM, é proposto o α-GMM para a classificação multiestilo de emoções e de condições de estresse. Este método foi inicialmente proposto para a tarefa de identificação de locutor [19]. Por meio de um fator de α, a capacidade de modelagem do GMM é expandida, o que é mais apropriado em condições de variações acústicas. A integração α generaliza a combinação linear adotada pelo GMM convencional. Para valores de α menores que -1, o classificador α-GMM enfatiza os maiores valores de probabilidade e atenua os menores. Uma vez que estados afetivos são variações acústicas introduzidas na voz em sua produção, entende-se que o α-GMM aumenta a performance de reconhecimento. Assim como o que foi mostrado em [19], outros trabalhos recentes demonstraram que o α-GMM consegue atingir resultados superiores ao clássico GMM [92, 93]. Por isso, resolveu-se adotá-lo como mais um classificador neste trabalho.

A partir de um modelo de estado afetivo λ_ε, composto por G densidades Gaussianas $b_g(\mathbf{x})$, $g = 1, ..., G$, a integração α das densidades é definida como [19]

$$p\left(\mathbf{x}|\lambda_\varepsilon\right) = C f_\alpha^{-1}\left\{\sum_{g=1}^{G} w_g f_\alpha\left[b_g(\mathbf{x})\right]\right\}, \tag{4.9}$$

em que w_g são os pesos das componentes Gaussianas, assim como apresentado na Equação 4.1, e C é uma constante de normalização. O termo $f_\alpha(\cdot)$ é dado por

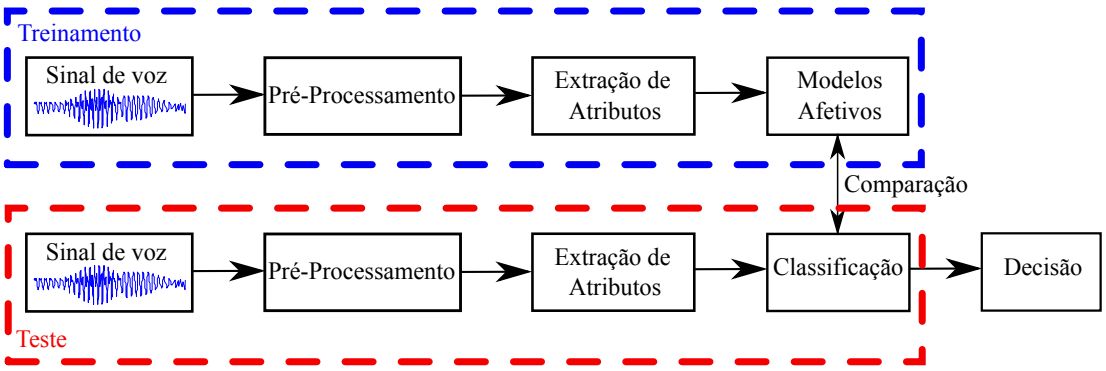


Figura 4.1 – Diagrama do sistema de classificação.

$$f_\alpha(x) = \begin{cases} \left(\frac{2}{1-\alpha}\right) x^{(1-\alpha)/2}, & \alpha \neq 1 \\ \log(x), & \alpha = 1. \end{cases} \tag{4.10}$$

Da Equação 4.10, a inversa de $f_\alpha(\cdot)$ pode ser calculada por

$$f_\alpha^{-1}(y) = \begin{cases} \left(\frac{1-\alpha}{2} y\right)^{\frac{2}{1-\alpha}}, & \alpha \neq 1 \\ \exp(y), & \alpha = 1. \end{cases} \tag{4.11}$$

Assim, o α-GMM de um estado emocional ou de uma condição de estresse pode ser reescrito como

$$p\left(\mathbf{x}|\lambda_\varepsilon\right) = C \left[\sum_{g=1}^{G} w_g b_g(\mathbf{x})^{\frac{1-\alpha}{2}}\right]^{\frac{2}{1-\alpha}}, \tag{4.12}$$

em que $\alpha = -1$ corresponde ao GMM convencional.

4.2 – Cenário dos Experimentos

O sistema de classificação utilizado nesta Tese está representado na Figura 4.1. Para ambas as etapas de treinamento e de teste, há os procedimentos de pré-processamento e extração de atributos. Os modelos afetivos formado a partir das características dos sinais acústicos são gerados no treinamento. Para cada sinal de voz na fase de teste, a matriz de atributos obtida é comparada com cada modelo no dicionário construído no treinamento. Assim, é decidido a qual estado afetivo pertence o sinal de entrada no teste. No contexto do classificador SVM, a procura pelo hiperplano ótimo é realizado o procedimento de *grid-search* para o *kernel* RBF, com os parâmetros de controle sendo examinados para c $\in$ (0, 10) e $\gamma \in$ (0, 1).

Para a classificação, é aplicado o procedimento de validação cruzada conhecido na literatura como LOSO (*Leave-One-Speaker-Out*) [6]. Neste método, a etapa de treinamento é realizada com todos os locutores do banco de dados, exceto aquele que será usado na etapa de teste. Isso

ocorre de forma que todos os locutores sejam utilizados tanto no treinamento quanto no teste. A etapa de treinamento foi conduzida com 32 s de cada estado afetivo, enquanto que os testes são empregados em segmentos de 800 ms dos sinais acústicos.

Na classificação multiestilo, o α-GMM é aplicado com cinco valores de α: -1 (GMM clássico), -2, -4, -6 e -8. Os modelos afetivos são compostos de 32 densidades Gaussianas com matrizes de covariância diagonais. O HMM é empregado com a topologia *left-to-right*, considerando 5 estados do modelo com uma mistura Gaussiana por estado, com o *software* HTK *toolkit* [94]. Para a técnica SVM, é utilizada a estratégia *one-versus-one* na classificação com a biblioteca LIBSVM [95].

4.2.1 – Bases Acústicas de Variações Afetivas

Para a proposta de um novo atributo para a classificação de variações acústicas afetivas não estacionárias, é importante que ele apresente robustez independente do idioma presente nos sinais. No contexto de bases acústicas de variações afetivas, um dos principais desafios ao longo dos anos é construir um banco de dados com estados afetivos captados da forma mais natural possível [4, 9, 96]. Os principais tipos de cenários utilizados na construção de bases acústicas com variações afetivas são três [96]: i) atuação, em que roteiros são seguidos de acordo com a pré-definição dos estados afetivos; ii) comportamento induzido, em que os locutores são provocados a sentir determinadas emoções; e iii) comportamento espontâneo, em que diálogos ocorrem de forma mais natural e não seguem roteiro nem indução de estados afetivos. Na proposta do vetor HHHC, além de os experimentos serem conduzidos com bases de dados gravadas em diferentes idiomas e lugares, elas compreendem pelo menos um desses três cenários.

Para atender às questões de diferentes idiomas e cenários de gravação, foram empregadas nesta Tese cinco bases acústicas. Quatro delas são utilizadas para analisar variações acústicas emocionais: *Berlin Database of Emotional Speech* (EMO-DB) [54], *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) [97], *Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression* (SEMAINE) [98] e *REmote COLaborative and Affective interactions* (RECOLA) [96]. No contexto de condições de estresse, é utilizada a base *Speech Under Simulated and Atual Stress* (SUSAS) [36] database. Cada uma das bases acústicas são descritas a seguir.

• EMO-DB

A base EMO-DB [54] foi desenvolvida na Universidade Técnica de Berlin, na Alemanha. Um total de 40 atores realizaram gravações em 7 emoções diferentes no idioma alemão: Raiva, Tédio, Nojo, Medo, Felicidade, Tristeza e Neutro. Estão catalogadas aproximadamente 800 gravações. Especialistas em análise perceptivo-auditiva selecionaram 10 destes atores, utilizando como critérios a naturalidade das gravações e o nível de reconhecimento auditivo

das emoções. De forma a balancear a base em relação a gênero, foram selecionados 5 homens e 5 mulheres, totalizando 494 sinais. A taxa de amostragem dos sinais inicialmente (quando gravados) era de 48.000 amostras/s. Depois os sinais passaram por um processo de subamostragem, em que os mesmos ficaram com uma taxa de 16.000 amostras/s. No Quadro 4.1, estão apresentadas as sentenças listadas na base EMO-DB.

Quadro 4.1 – Sentenças Listadas na Base EMO-DB.

Der Lappen liegt auf dem Eisschrank.
Das will sie am Mittwoch abgeben.
Heute abend könnte ich es ihm sagen.
Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.
In sieben Stunden wird es soweit sein.
Was sind denn das für Tüten, die da unter dem Tisch stehen?
Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.
An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.
Ich will das eben wegbringen und dann mit Karl was trinken gehen.
Die wird auf dem Platz sein, wo wir sie immer hinlegen.

Das emoções catalogadas na EMO-DB, nesta Tese foram utilizadas as seguintes: Raiva, Felicidade, Tédio, Tristeza e Neutro. Após a seleção dos trechos sonoros de cada sinal acústico, chegou-se a um total de 40 segundos por estado emocional considerado nos experimentos.

• IEMOCAP

A base IEMOCAP [97] foi desenvolvida no *Speech Analysis and Interpretation Laboratory* (SAIL) da Universidade do Sul da Califórnia, nos Estados Unidos. A coleta dos sinais acústicos foi realizada em interações dois a dois, com 10 atores (5 homens e 5 mulheres). Desse grupo de atores, 7 são profissionais e 3 eram alunos *Senior* no Departamento de Drama da Universidade do Sul da Califórnia na época da formação da base de dados audiovisuais. Nas interações entre os atores, as conversações consistiam de dois tipos de cenários: situações hipotéticas (seguindo roteiros) e diálogos espontâneos realizados de forma improvisada entre eles. As gravações dos sinais acústicos, no idioma inglês, utilizaram uma taxa de 48.000 amostras/s.

Os estados emocionais considerados neste trabalho de Tese nas análises com a base IEMOCAP foram: Raiva, Felicidade, Neutro e Tristeza. Para cada um desses 4 estilos, foi utilizado 10 minutos de trechos sonoros dos sinais de voz.

• SEMAINE

A base SEMAINE [98] foi desenvolvida a partir de uma cooperação de uniersidades da Inglaterra, da Holanda e da Alemanha. Na coleta dos dados audiovisuais, participaram 150 estudantes de graduação e pós gradução de oito diferente países. Para a indução dos estados emocionais, foi utilizado o cenário SAL (*Sensitive Artificial Listener*) e o idioma falado foi o inglês. Neste cenário, o participante é convidado a falar sobre tópicos que são

emocionalmente significantes para eles, que são provocados a expressar fortemente as emoções por meio da inclusão de palavras-chave no diálogo. Nas interações, realizadas dois a dois, foram considerados um "usuário" (humano) e um "operador" (que pode ser um humano ou uma máquina). Neste tipo de interação, o controle do conteúdo da conversação ficava por conta do operador. Os sinais acústicos foram gravados a uma taxa de 48.000 amostras/s. Julgadores dividiram as emoções nos trechos das gravações em cinco dimensões: valência, ativação, potência, antecipação/expectativa, e intensidade. Além disso, as emoções foram divididas em 27 estilos, entre os quais estão Raiva e Tristeza, além de outros comportamentos, como demonstração de solidariedade e antagonismo.

Neste trabalho de Tese, foi considerado as gravações de 10 participantes (5 homens e 5 mulheres). No que diz respeito às emoções analisadas, além de Raiva e Tristeza, Diversão e Felicidade foram incluídas para a avaliação no contexto multiestilo. Para cada estado emocional, foi utilizado 90 segundos de trechos sonoros dos sinais de voz.

• RECOLA

A base RECOLA [96] foi desenvolvida na Suíça, no Departamento de Psicologia da Universidade de Fribourg. Um total de 46 participantes foram submetidos a interações dois a dois. Os sinais acústicos foram captados no idioma Francês, havendo 33 participantes nativos da língua francesa, 8 italianos, 4 alemães e 1 português. Antes de haver a interação entre os participantes, eles foram submetidos a um questionário de auto-avaliação emocional conhecido como SAM (*Self-Assessment Manikin*), que está relacionado com a valência das emoções. Os aplicadores do questionário decidiram quais participantes iriam ser induzidos[1] com humor positivo ou negativo, de acordo com o SAM. Então, as interações aconteciam de forma remota enquanto os sinais acústicos e biológicos eram capturados. Após a obtenção das gravações, julgadores analizaram 5 minutos de cada interação dois a dois a fim de mapear os sinais acústicos de acordo com duas dimensões: valência e ativação. A taxa de amostragem utilizada nas gravações é de 44.100 amostras/s.

Nos experimentos realizados neste trabalho de Tese, apenas foi considerada a dimensão da ativação. Para um balanceamento na análise, foram selecionados 5 locutores homens e 5 mulheres. No caso desta base acústica, as emoções estão separadas em alta e baixa ativação, sendo considerados 10 minutos de trechos sonoros de cada estado de ativação das emoções.

• SUSAS

A base SUSAS [99] foi desenvolvida na Universidade Duke, nos Estados Unidos, como forma de analisar variações acústicas causadas por diferentes níveis de estresse. Esta base é composta por 3.593 sinais captados a uma taxa de 8.000 amostras/s, em condições reais de

[1] A indução do humor na base RECOLA foi realizada por meio de apresentação de vídeo clipes voltados a humor positivo ou negativo, a depender do apurado de cada participante na auto-avaliação do SAM.

estresse e medo. Para isto, os sinais foram obtidos com 7 locutores (4 homens e 3 mulheres), submetidos a duas situações distintas: montanha-russa e queda livre. A base SUSAS aborda as situações de alto estresse, médio estresse e grito, além do estado neutro. Diferentemente da base EMO-DB, cujas locuções correspondem a sentenças de diferentes tamanhos, a base SUSAS possui 35 comandos de curta duração na língua inglesa, que estão elencados no Quadro 4.2.

Quadro 4.2 – Comandos Listados na Base SUSAS.

break	enter	help	on	strafe
change	fifty	histogram	out	ten
degree	fix	destination	point	thirty
hot	freeze	mark	six	three
east	gain	nav	south	white
eight	go	no	stand	wide
eighty	hello	oh	steer	zero

As 4 condições de estresse catalogadas na base SUSAS são analisadas neste trabalho de Tese. De forma a balancear os experimentos, a duração dos sinais acústicos de cada estado afetivo foi mantida em 146 segundos, após a seleção dos trechos sonoros dos sinais de voz.

4.2.2 – Atributos Acústicos Utilizados na Classificação

Para a análise do potencial de classificação do atributo proposto (HHHC), a metodologia de extração segue o que foi apresentado na Seção 3.2.3. O vetor HHHC utilizado tem dimensão igual a 6. Para a análise com EEMD, 11 níveis de ruído Gaussiano branco foram utilizados nos experimentos, considerando o desvio-padrão do ruído: $0,005$, $0,01$, $0,02$, $0,03$, $0,04$, $0,05$, $0,06$, $0,07$, $0,08$, $0,09$ and $0,1$. Na fusão com o INS, ele é calculado dentro de cada IMF considerando as seguintes escalas de observação (T_h/T): $0,0015$, $0,025$, $0,05$, $0,1$, $0,15$, $0,2$, $0,25$, $0,3$, $0,4$ e $0,5$.

Para fins comparativos, são utilizados três atributos encontrados na literatura de classificação de emoções e condições de estresse: o vetor pH (atributo acústico da fonte de excitação), os coeficientes MFCC (atributo acústico do trato vocal) e o atributo TEO (atributo não linear). Os vetores pH são obtidos em quadros de 50 ms, obtidos a cada 10 ms usando a transformada wavelet com filtros de Daubechies com 12 coeficientes (escalas 2-12). Em relação ao MFCC, são obtidos 12 coeficientes dos sinais de voz em segmentos de 25 ms, com uma taxa de quadro de 10 ms. Para o atributo TEO, vetores com 16 coeficientes são obtidos de quadros de 75 ms dos sinais acústicos, com 50% de sobreposição.

4.3 – Resultados

Os resultados obtidos do procedimento de classificação são apresentados nesta Seção. Nos experimentos, foi observado que com baixos níveis de ruído (desvio padrão entre $0,005$ e $0,02$) na decomposição com EEMD, o vetor HHHC atinge seus melhores resultados. Além da

coomparação do HHHC com outros atributos, são mostrados os resultados de 4 classificadores clássicos, em comparação com o proposto α-GMM.

4.3.1 – Classificação com a base EMO-DB

Os resultados obtidos da classificação utilizando HHHC e os demais atributos considerados com a base EMO-DB são apresentados na Tabela 4.1. No que diz respeito aos classificadores estocásticos clássicos, o GMM atinge resultados superiores ao HMM. O vetor HHHC obtém uma acurácia média de $76,4\%$ com GMM e $74,6\%$ com HMM. Esta diferença de aproximadamente 1 e 2 pontos percentuais (p.p.) entre GMM e HMM ocorre para os demais atributos. Ainda no contexto do classificador GMM, pode ser observado que a acurácia média do vetor HHHC supera o vetor pH em 12,4 p.p., 19,6 p.p. os coeficientes MFCC e 25,6 p.p. o atributo TEO. Ao considerar as taxas de acerto de cada estado emocional individualmente, pode ser observado a vantagem do atributo HHHC. Com GMM e HMM, todas as emoções são classificadas com pelo menos 67% de acerto. Em relação ao SVM, as taxas de acerto não superam aquelas obtidas pelos métodos estocásticos. O vetor HHHC obteve uma acurácia média de $64,2\%$, o que representa mais de 10 p.p. abaixo do GMM. Todavia, com o SVM foi obtida uma taxa de identificação de pelo menos 51% para todos os estados emocionais, o que não ocorre para os atributos acústicos comparativos. Em relação ao vetor pH, o HHHC atinge 36 p.p. a mais na classificação da emoção tédio, utilizando GMM. Para esta mesma emoção ocorre a maior diferença de acurácia entre HHHC e MFCC: 38 p.p. utilizando GMM. Na identificação dos estados emocionais Raiva e Neutro, o vetor HHHC obtém as taxas de acertos maiores que aquelas obtidas com o atributo TEO: diferença de 51 p.p. e 57 p.p., respectivamente.

Os resultados obtidos com o classificador proposto α-GMM são apresentados na Figura 4.2. Esta abordagem estocástica atinge resultados superiores do que aqueles obtidos pelos métodos clássicos. A maior taxa de acurácia média para os atributos comparativos foi obtida com pH ($65,4\%$) com $\alpha = -2$. A maior taxa de acerto para o vetor HHHC ($79,2\%$) é atingida com três valores de α (-4, -6 and -8). Ainda, o HHHC supera em 15,6 p.p. a acurácia média obtida pelo MFCC ($63,6\%$) e em 26,4 p.p. o que foi obtido pelo atributo TEO (52.8%). Em relação à análise da emoções individualmente, a performance do HHHC permite taxas de acerto acima de 60% para todas as analisadas com α-GMM. Para Raiva, HHHC supera TEO em 53 p.p., ambos considerando $\alpha = -8$. Ainda em relação ao TEO, HHHC atinge 90% de acerto para o estado Neutro ($\alpha = -8$) contra 47% ($\alpha = -4$). Na classificação da emoção Felicidade, a maior diferença nas principais taxas de acerto entre HHHC e pH foi obtida com 75% (com $\alpha = -6$) para HHHC e com 67% (com $\alpha = -4$) para pH. Na comparação com MFCC pode ser observado que há a maior diferença de acurácia em relação a HHHC no contexto das emoções Tédio e Tristeza. Nestes casos, o atributo proposto supera MFCC em 40 p.p. (75% com HHHC e 35% com MFCC, considerando $\alpha = -6$) e 21 p.p. (82% para HHHC com $\alpha = -2$ e 68% para MFCC com $\alpha = -8$), respectivamente.

Tabela 4.1 – Taxas de acurácia (%) de 5 estados emocionais considerando os classificadores GMM, HMM e SVM para a base EMO-DB.

	Emoção	Emoção Classificada com GMM					Emoção Classificada com HMM					Emoção Classificada com SVM				
	Real	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.
HHHC	Raiva	**78**	22	0	0	0	**76**	24	0	0	0	**72**	28	0	0	0
	Felicidade	29	**71**	0	0	0	33	**67**	0	0	0	37	**63**	0	0	0
	Neutro	0	0	**82**	18	0	0	0	**81**	19	0	0	0	**64**	34	2
	Tédio	0	0	15	**69**	16	0	0	15	**68**	17	0	0	20	**51**	29
	Tristeza	0	0	0	18	**82**	0	0	0	19	**81**	0	0	0	29	**71**
		Taxa de classificação média: **76,4**					Taxa de classificação média: **74,6**					Taxa de classificação média: **64,2**				
pH	Emoção	Emoção Classificada com GMM					Emoção Classificada com HMM					Emoção Classificada com SVM				
	Real	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.
	Raiva	**80**	20	0	0	0	**78**	22	0	0	0	**69**	30	1	0	0
	Felicidade	30	**65**	5	0	0	32	**64**	4	0	0	35	**57**	8	0	0
	Neutro	0	6	**67**	17	10	0	6	**64**	20	10	0	8	**56**	24	12
	Tédio	0	11	24	**33**	32	0	5	31	**33**	31	0	9	28	**27**	36
	Tristeza	0	3	7	15	**75**	0	3	8	15	**74**	0	2	10	20	**68**
		Taxa de classificação média: **64,0**					Taxa de classificação média: **62,6**					Taxa de classificação média: **55,4**				
MFCC	Emoção	Emoção Classificada com GMM					Emoção Classificada com HMM					Emoção Classificada com SVM				
	Real	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.
	Raiva	**73**	25	2	0	0	**74**	24	2	0	0	**63**	30	7	0	0
	Felicidade	25	**71**	4	0	0	25	**70**	5	0	0	27	**65**	8	0	0
	Neutro	0	12	**49**	28	11	0	19	**48**	23	10	0	20	**43**	25	12
	Tédio	0	6	30	**31**	33	0	8	34	**28**	30	0	11	37	**19**	33
	Tristeza	0	4	10	26	**60**	0	5	11	25	**59**	0	12	24	35	**29**
		Taxa de classificação média: **56,8**					Taxa de classificação média: **55,8**					Taxa de classificação média: **43,8**				
TEO	Emoção	Emoção Classificada com GMM					Emoção Classificada com HMM					Emoção Classificada com SVM				
	Real	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.
	Raiva	**27**	51	22	0	0	**28**	52	20	0	0	**20**	56	24	0	0
	Felicidade	33	**63**	4	0	0	31	**59**	5	5	0	30	**55**	10	5	0
	Neutro	10	24	**25**	37	4	10	34	**24**	32	0	13	36	**20**	31	0
	Tédio	1	7	20	**53**	19	3	6	26	**51**	14	4	7	27	**47**	15
	Tristeza	0	0	0	14	**86**	4	0	6	15	**75**	7	7	0	17	**69**
		Taxa de classificação média: **50,8**					Taxa de classificação média: **47,4**					Taxa de classificação média: **42,2**				

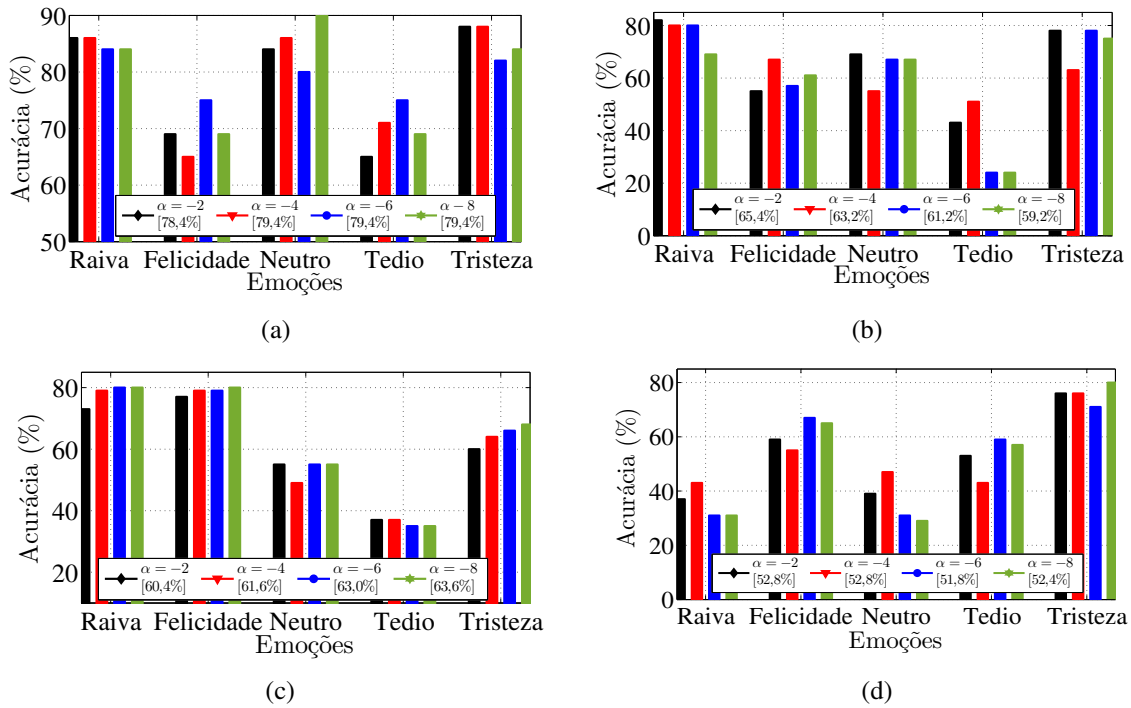


Figura 4.2 – Acurácia obtida da classificação utilizando α-GMM com a base EMO-DB, para os seguintes atributos: (a) HHHC; (b) pH; (c) MFCC; (d) TEO.

4.3.2 – Classificação com a base IEMOCAP

Na Tabela 4.2 estão os resultados pertinentes à classificação realizada com a base IEMOCAP. O classificador GMM atinge taxas de acerto superiores àquelas obtidas com o HMM e SVM. No contexto dos classificadores estocásticos clássicos, a maior acurácia média é obtida com o HHHC, com GMM ($54,5\%$), enquanto que o mesmo atributo com HMM atinge $52,0\%$ de acerto e $42,3\%$ com SVM. Este resultado do vetor HHHC com classificador GMM é obtido contra $50,8\%$ com o pH, $47,8\%$ obtido pelo MFCC e $42,0\%$ obtido pelo TEO. A maior diferença entre as taxas de acerto do HHHC e dos atributos comparativos é no caso HHHC *versus* TEO, considerando as emoções Raiva, Felicidade e Neutro (18 p.p., 16 p.p. e 20 p.p. de diferença, respectivamente). Com GMM, HHHC supera os 49% de acerto para todas as emoções presentes nos experimentos com a IEMOCAP. O classificador SVM atinge resultados inferiores àqueles obtidos pelo GMM e HMM. Em relação aos atributos comparativos utilizando SVM, o HHHC supera o pH em 4,5 p.p.. NO contexto do MFCC, o HHHC o supera em 6,3 p.p., enquanto que atinge 9 p.p. a mais que TEO. Na base IEMOCAP, embora os valores das taxas de acerto tenham sido menores que aqueles obtidos para a base EMO-DB, o atributo HHHC atinge os melhores resultados independente do classificador clássico considerados nestes experimentos.

Tabela 4.2 – Taxas de acurácia (%) de 4 estados emocionais considerando os classificadores GMM, HMM e SVM para a base IEMOCAP.

	Emoção	Emoção Classificada com GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.
HHHC	Raiva	**57**	27	12	4	**55**	28	12	5	**49**	31	14	6
	Felicidade	30	**49**	19	2	31	**45**	19	5	30	**35**	28	7
	Neutro	10	15	**54**	21	10	15	**54**	21	15	20	**39**	26
	Tristeza	7	11	24	**58**	7	12	27	**54**	7	14	33	**46**
		Taxa de classificação média: **54,5**				Taxa de classificação média: **52,0**				Taxa de classificação média: **42,3**			
pH	Emoção	Emoção Classificada com GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.
	Raiva	**59**	24	13	4	**57**	26	13	4	**49**	30	15	6
	Felicidade	32	**43**	17	8	33	**42**	17	8	29	**30**	26	15
	Neutro	12	14	**51**	23	12	15	**49**	24	17	24	**32**	27
	Tristeza	6	16	28	**50**	10	14	27	**49**	12	15	33	**40**
		Taxa de classificação média: **50,8**				Taxa de classificação média: **49,3**				Taxa de classificação média: **37,8**			
MFCC	Emoção	Emoção Classificada com GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.
	Raiva	**53**	18	16	13	**50**	19	18	13	**40**	22	23	15
	Felicidade	32	**43**	22	5	30	**37**	22	11	32	**32**	24	12
	Neutro	16	13	**45**	26	16	12	**44**	28	18	15	**31**	36
	Tristeza	8	12	28	**52**	10	12	28	**50**	13	15	31	**41**
		Taxa de classificação média: **47,8**				Taxa de classificação média: **45,3**				Taxa de classificação média: **36,0**			
TEO	Emoção	Emoção Classificada com GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.
	Raiva	**39**	26	23	12	**37**	26	25	12	**27**	30	29	14
	Felicidade	33	**33**	22	12	35	**31**	22	12	37	**25**	24	14
	Neutro	8	25	**34**	33	8	25	**33**	34	9	27	**26**	38
	Tristeza	7	5	22	**62**	9	8	24	**59**	9	9	27	**55**
		Taxa de classificação média: **42,0**				Taxa de classificação média: **40,0**				Taxa de classificação média: **33,3**			

No contexto da classificação com α-GMM, os resultados são apresentados na Figura 4.3. Como pode ser observado, o classificador proposto obtém valores de acurácia mais altos que aqueles obtidos com os classificadores clássicos. Note que apenas o vetor HHHC proporciona ao classificador uma taxa de acerto média de mais 60%. Isto ocorre para $\alpha = -8$. Em

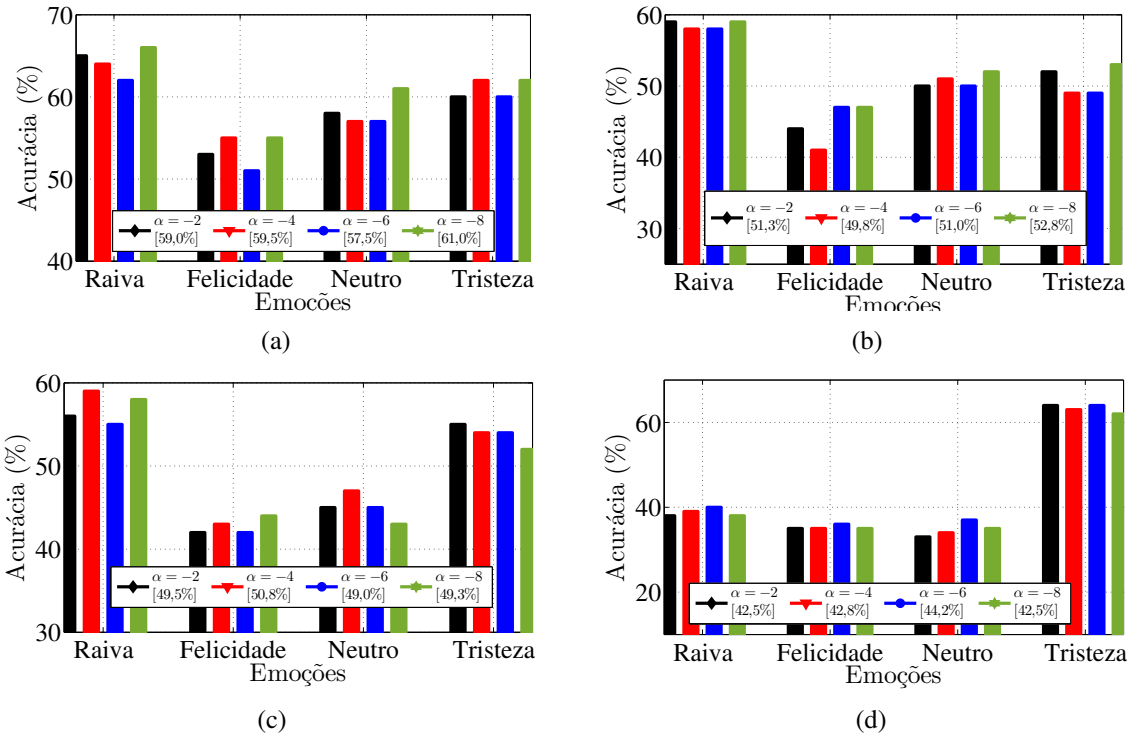


Figura 4.3 – Acurácia obtida da classificação utilizando α-GMM com a base IEMOCAP, para os seguintes atributos: (a) HHHC; (b) pH; (c) MFCC; (d) TEO.

comparação aos demais atributos, HHHC obteve uma acurácia média 8 p.p. acima daquela obtida com o vetor pH ($\alpha = -8$), 10 p.p. acima do MFCC ($\alpha = -4$) e 15 p.p. a mais que o atributo TEO ($\alpha = -6$). Para cada estado emocional analisado no contexto do HHHC, o classificador α-GMM obtém mais de $50,0\%$ de taxa de acerto. Em relação às emoções Raiva, Felicidade e Neutro, a maior diferença nas taxas de acerto diz respeito ao caso HHHC *versus* TEO. Nestes casos com $\alpha = -8$, por exemplo, o atributo proposto nesta Tese supera TEO em 28 p.p., 20 p.p. e 26 p.p., respectivamente. Com este mesmo valor de α, em relação à Tristeza, HHHC atinge 9 p.p. acima do obtido com o vetor pH.

4.3.3 – Classificação com a base SEMAINE

Para a base SEMAINE, os resultados do processo de classificação das variações acústicas afetivas estão apresentados na Tabela 4.3. Os valores de acurácia referem-se aos classificadores estocásticos clássicos (GMM e HMM) e ao SVM. Assim como ocorreu nos casos das bases EMO-DB e IEMOCAP, o GMM tem desempenho melhor que o HMM no contexto da SEMAINE. Enquanto o HHHC atinge uma acurácia média de $51,3\%$ com GMM, seu desempenho com HMM chega a $48,8\%$. Em relação aos atributos acústicos comparativos e considerando o GMM, o vetor pH obtém a maior taxa de acerto média ($47,3\%$). Porém, este resultado ainda é 4 p.p. abaixo do HHHC. Quanto a MFCC e TEO, o vetor HHHC os supera em aproximadamente 5 p.p. e 14 p.p., respectivamente. No contexto dos estados emocionais

analisados individualmente, o HHHC supera o atributo TEO em 16 p.p., enquanto a diferença entre estes dois atributos para Felicidade e Diversão foi de 19 p.p. em ambos os casos. Note que as emoções que mais se confundem são Felicidade e Diversão, o que é esperado pelo fato de se tratarem de dois comportamentos semelhantes. Apesar disso, o HHHC consegue uma taxa de acerto acima de 50% para ambas. Na classificação com SVM, os resultados são inferiores aqueles obtidos com GMM e HMM. Em um comparativo entre GMM e SVM, por exemplo, o classificador estocástico atinge uma acurácia média 8,5 p.p. superior, considerando o atributo HHHC. Mesmo com resultados inferiores, com SVM o HHHC supera os atributos comparativos. O vetor pH, entre eles, foi o que obteve a maior acurácia média $(39,3\%)$, seguido por MFCC $(36,3\%)$ e TEO $(27,8\%)$.

Tabela 4.3 – Taxas de acurácia (%) de 4 estados emocionais considerando os classificadores GMM, HMM e SVM para a base SEMAINE.

	Emoção	Emoção Classificada com GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
HHHC	Real	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.
	Raiva	**46**	25	22	7	**45**	26	22	7	**39**	28	24	9
	Felicidade	15	**52**	28	5	17	**50**	28	5	20	**43**	32	5
	Diversão	14	29	**50**	7	14	29	**48**	9	16	32	**43**	9
	Tristeza	8	16	19	**57**	8	18	22	**52**	9	20	25	**46**
		Taxa de classificação média: **51,3**				Taxa de classificação média: **48,8**				Taxa de classificação média: **42,8**			
pH	Emoção	Emoção Classificada com GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.
	Raiva	**45**	25	22	8	**45**	25	22	8	**38**	29	25	8
	Felicidade	20	**48**	29	3	19	**47**	29	5	22	**40**	33	5
	Diversão	16	28	**47**	9	16	28	**45**	11	18	30	**39**	13
	Tristeza	8	18	25	**49**	8	18	27	**47**	9	20	31	**40**
		Taxa de classificação média: **47,3**				Taxa de classificação média: **46,0**				Taxa de classificação média: **39,3**			
MFCC	Emoção	Emoção Classificada com GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.
	Raiva	**40**	32	18	10	**38**	31	17	14	**30**	34	20	16
	Felicidade	19	**52**	25	4	19	**49**	28	4	21	**41**	33	5
	Diversão	16	31	**44**	9	16	31	**42**	11	18	34	**35**	13
	Tristeza	10	13	19	**51**	10	13	30	**47**	11	15	35	**39**
		Taxa de classificação média: **46,8**				Taxa de classificação média: **44,0**				Taxa de classificação média: **36,3**			
TEO	Emoção	Emoção Classificada com GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.
	Raiva	**30**	24	24	22	**28**	26	24	22	**18**	30	28	24
	Felicidade	32	**33**	30	5	30	**31**	30	9	33	**22**	35	10
	Diversão	20	27	**31**	22	20	27	**31**	22	21	29	**24**	26
	Tristeza	3	18	20	**57**	3	18	24	**55**	3	21	29	**47**
		Taxa de classificação média: **37,8**				Taxa de classificação média: **36,2**				Taxa de classificação média: **27,8**			

No contexto da classificação com α-GMM, os resultados estão na Figura 4.4. Note que os resultados são superiores aqueles obtidos com os classificadores clássicos. A maior taxa de acurácia média para o HHHC foi de $54,5\%$ usando $\alpha = -6$, o que representa um aumento de 3,2 p.p. em relação ao GMM (melhor resultado entre os classificadores clássicos). Estes resultados são obtidos contra $50,8\%$ ($\alpha = -4$) obtido pelo vetor pH, $49,0\%$ ($\alpha = -6$) obtido pelo MFCC, e $40,8\%$ ($\alpha = -8$) com o atributo TEO. Em relação a estes atributos comparativos, os resultados com α-GMM atinge, por exemplo, aproximadamente 4 p.p. a mais que o HMM e 10 p.p. a mais que o que foi obtido com HMM. Assim como observado no caso do GMM, com o α-GMM emoções similares como Felicidade e Diversão são classificadas com mais de $50,0\%$ de precisão, utilizando o atributo HHHC. Em uma avaliação individual de cada estado emocional, verifica-se que para os atributos HHHC e MFCC, todas as emoções são melhor

classificadas com valor de α igual a -6. Para o vetor pH, as maiores taxa de acerto estão concentradas entre os valores de α iguais a -4 e -6. Em relação ao atributo TEO, as maiores taxas de acerto para as emoções estão entre $-8 \leq \alpha \leq -6$. Para Raiva, a maior diferença observada ocorre entre HHHC e TEO (tanto para $\alpha = -2$ quanto para $\alpha = -6$), em que o atributo proposto supera TEO em 17 p.p. na taxa de acurácia média. Esta diferença chega a 27 p.p. para a emoção Felicidade considerando,para ambos os atributos, $\alpha = -6$. Com este mesmo valor de α, a diferença entre HHHC e TEO é de 16 p.p. para a emoção Diversão. Para a emoção Tristeza, HHHC supera em 5 p.p. os atributos pH e MFCC, também considerando $\alpha = -6$.

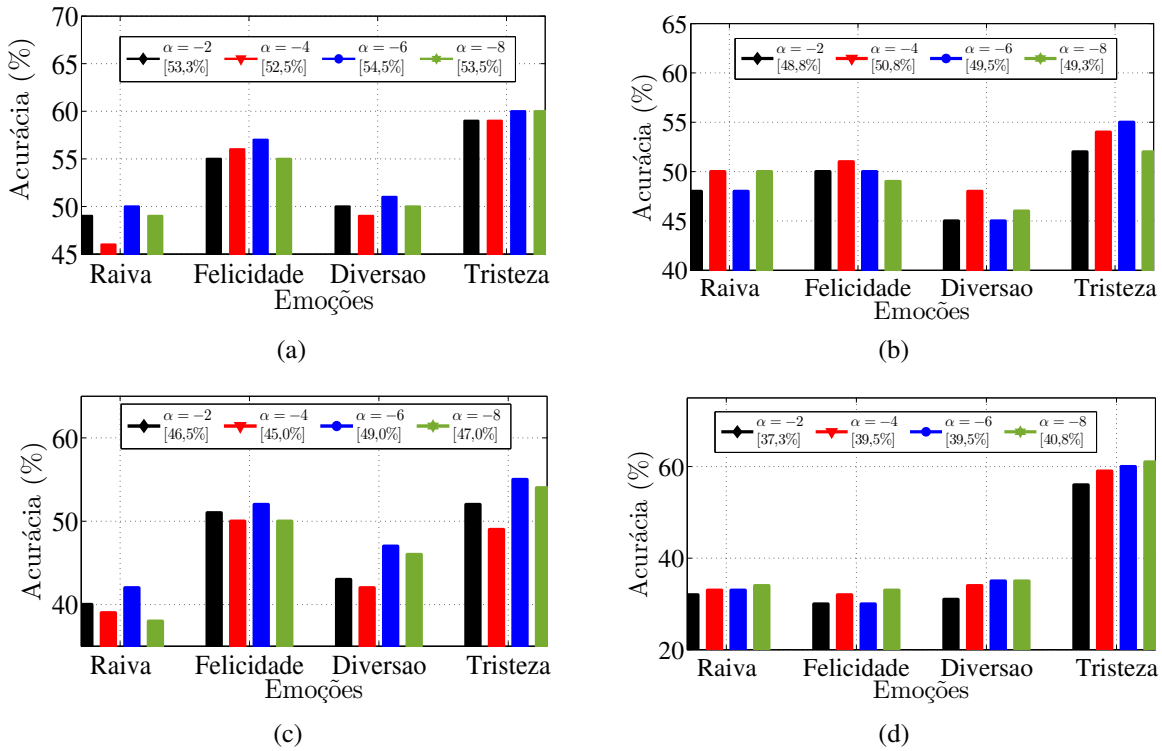


Figura 4.4 – Acurácia obtida da classificação utilizando α-GMM com a base SEMAINE, para os seguintes atributos: (a) HHHC; (b) pH; (c) MFCC; (d) TEO.

4.3.4 – Classificação com a base RECOLA

Os valores de acurácia obtidos para a base RECOLA são apresentados na Tabela 4.4. Diferente das demais bases acústicas com variações emocionais, a RECOLA tem seus estados emocionais agrupados de acordo com o nível de ativação (alta ou baixa). Em relação aos classificadores estocásticos clássicos, o GMM obtém desempenho melhor que o HMM. Enquanto que com GMM o HHHC atinge uma acurácia média de $58,5\%$, com HMM sua taxa de acerto é de $55,0\%$. No contexto do GMM, o HHHC sozinho supera pH, MFCC e TEO em 5 p.p., 13,5 p.p. e 14 p.p., respectivamente. Apenas com HHHC e pH, as taxas de acerto individuais superam os $50,0\%$. No caso do atributo proposto, estas taxas passam de $55,0\%$.

O SVM tem desempenho inferior aos classificadores GMM e HMM. Por exemplo, no caso do HHHC, GMM tem performance 9 p.p. superior ao SVM. Em relação ao atributo MFCC, o GMM atinge 6 p.p. acima do SVM. No caso do SVM, apenas o HHHC+INS proporciona uma classificação acima de $48,0\%$ de ambos os estados afetivos. Para todos os classificadores clássicos, o atributo proposto supera os atributos comparativos.

Tabela 4.4 – Taxas de acurácia (%) em relação ao nível de ativação dos estados emocionais considerando os classificadores GMM, HMM e SVM para a base RECOLA.

	Emoção	Classificação com GMM		Classificação com HMM		Classificação com SVM	
HHHC	Real	Alta	Baixa	Alta	Baixa	Alta	Baixa
	Alta ativação	**58**	42	**56**	44	**49**	51
	Baixa ativação	41	**59**	46	**54**	50	**50**
		Taxa de classificação média: **58,5**		Taxa de classificação média: **55,0**		Taxa de classificação média: **49,5**	
pH	Emoção	Classificação com GMM		Classificação com HMM		Classificação com SVM	
	Real	Alta	Baixa	Alta	Baixa	Alta	Baixa
	Alta ativação	**54**	46	**52**	48	**46**	54
	Baixa ativação	47	**53**	50	**50**	48	**52**
		Taxa de classificação média: **53,5**		Taxa de classificação média: **51,0**		Taxa de classificação média: **51,0**	
MFCC	Emoção	Classificação com GMM		Classificação com HMM		Classificação com SVM	
	Real	Alta	Baixa	Alta	Baixa	Alta	Baixa
	Alta ativação	**46**	54	**45**	55	**40**	60
	Baixa ativação	56	**44**	53	**47**	62	**38**
		Taxa de classificação média: **45,0**		Taxa de classificação média: **46,0**		Taxa de classificação média: **39,0**	
TEO	Emoção	Classificação com GMM		Classificação com HMM		Classificação com SVM	
	Real	Alta	Baixa	Alta	Baixa	Alta	Baixa
	Alta ativação	**46**	54	**43**	57	**37**	63
	Baixa ativação	57	**43**	60	**40**	64	**36**
		Taxa de classificação média: **44,5**		Taxa de classificação média: **41,5**		Taxa de classificação média: **36,5**	

Na Figura 4.5 são apresentados os resultados obtidos com o classificador α-GMM. Este método obtém melhor desempenho que os clássicos, para todos os atributos considerados nos experimentos. O vetor HHHC atinge a maior taxa de acurácia média utilizando $\alpha = -6$. Entre os demais atributos, a maior taxa média de acerto ocorre com pH ($59,5\%$ com $\alpha = -4$).Também com $\alpha = -4$, MFCC atinge seu melhor resultado, com $54,5\%$ de acerto. Para o atributo TEO, com o valor de $\alpha = -2$ é atingido o seu melhor desempenho, com uma taxa de acurácia média igual a $49,5\%$. Para todos os valores de α analisados, o vetor HHHC proporciona ao classificador uma performance superior a $58,0\%$ na classificação de ambos os estados afetivos.

4.3.5 – Classificação com a base SUSAS

Na Tabela 4.11, estão apresentados os resultados da classificação realizada com os atributos acústicos extraídos da base SUSAS. Em relação aos classificados GMM e HMM, o primeiro atinge resultados superiores. Embora a base SUSAS trate de condições de estresse (o que é uma abordagem diferente do contexto de emoções), pode ser observado que o vetor HHHC proporciona aos classificadores desempenho superior aos atributos comparativos, assim como ocorreu para as demais bases. A condição de Grito é a que mais se distingue das demais. Com o vetor HHHC, ela chega a ser identificada com 100%. O vetor HHHC supera, em termos de acurácia média, os atributos pH, MFCC e TEO em 11,8 p.p., 13,5 p.p. e 14,3 p.p., respectivamente. No contexto do SVM, os resultados são inferiores aqueles atingidos por GMM

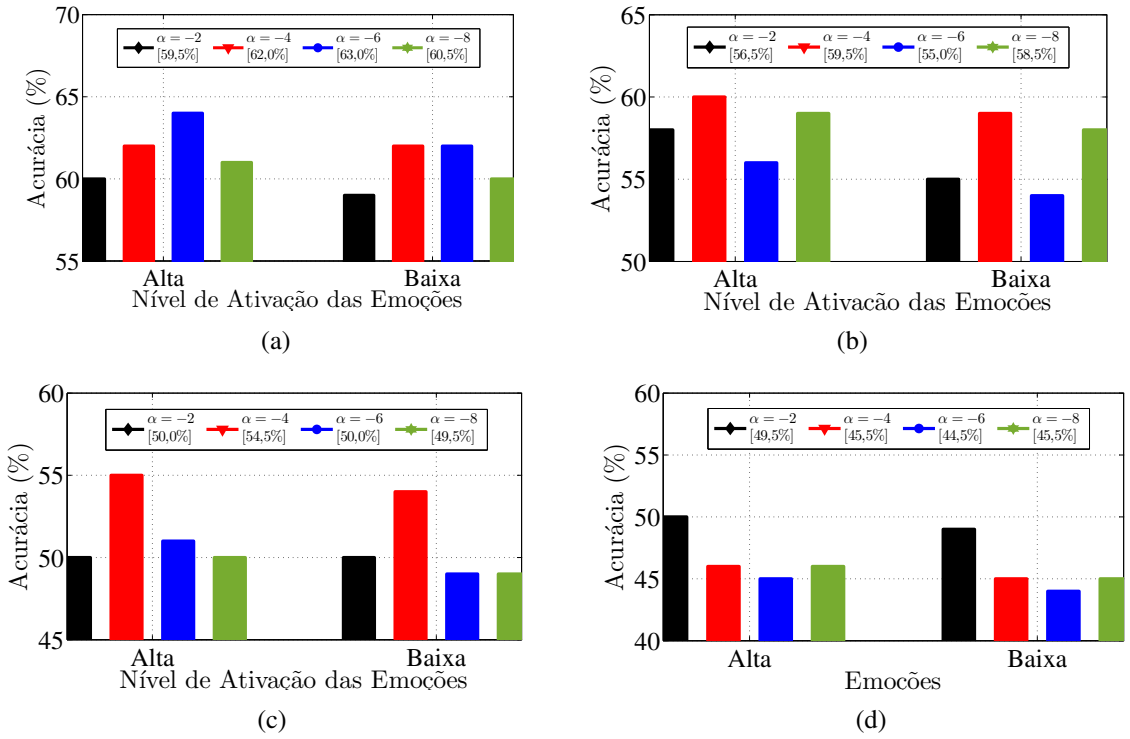


Figura 4.5 – Acurácia obtida da classificação utilizando α-GMM com a base RECOLA, para os seguintes atributos: (a) HHHC; (b) pH; (c) MFCC; (d) TEO.

e HMM. Com o classificador SVM, o atributo proposto é o único que chega a uma taxa de acerto de mais de $50,0\%$ para todas as condições de estresse consideradas nos experimentos. A maior diferença para os estados de Alto estresse e Grito, por exemplo, ocorre entre HHHC e pH, em que o primeiro supera o segundo em 10 p.p. e 3 p.p., respectivamente.

No que diz respeito ao classificador α-GMM, seus resultados estão apresentados na Figura 4.6, em que pode ser observado que as taxas de acerto são superiores àquelas obtidos pelos classificadores clássicos. A diferença na acurácia média entre α-GMM e os demais classificadores chega a mais de 4 p.p.. O HHHC atinge uma acurácia média de $76,3\%$ com $\alpha = -2$ e $76,0\%$ com os demais valores de $alpha$. Isto representa 16 p.p. maior que o valor de acurácia média obtido pelo atributo TEO ($60,0\%$ com $\alpha = -4$). Além disso, HHHC supera em 14 p.p. e 11.5 p.p. os atributos MFCC ($62,0\%$ com $\alpha = -4$) e pH ($64,5\%$ com $\alpha = -8$), respectivamente. Note que, para os atributos comparativos, Médio estresse é classificado com uma acurácia abaixo de $40,0\%$, enquanto que esta variação acústica é classificada com $72,0\%$ ($\alpha = -4$) de acerto com HHHC. Assim, como ocorreu com GMM, o estado de Grito é classificado com 100% de acerto com o α-GMM.

4.3.6 – Principais Resultados dos Atributos

Na Tabela 4.6 é mostrado um resumo dos melhores resultados de classificação obtidos com todos os atributos extraídos das bases de dados utilizadas nos experimentos. Em todos os casos,

Tabela 4.5 – Taxas de acurácia (%) de 4 condições de estresse considerando os classificadores GMM, HMM e SVM para a base SUSAS.

	Condição	Cond. Classificada com GMM				Cond. Classificada com HMM				Cond. Classificada com SVM			
HHHC	Real	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.
	Neutro	**71**	21	8	0	**69**	21	8	2	**63**	24	10	3
	Médio	23	**59**	18	0	23	**59**	18	0	27	**51**	22	0
	Alto	5	17	**61**	17	5	17	**60**	18	7	20	**50**	23
	Grito	0	0	0	**100**	0	0	2	**98**	0	0	10	**90**
		Taxa de classificação média: **72,8**				Taxa de classificação média: **71,5**				Taxa de classificação média: **63,5**			
pH	Condição	Cond. Classificada com GMM				Cond. Classificada com HMM				Cond. Classificada com SVM			
	Real	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.
	Neutro	**76**	19	5	0	**75**	17	8	0	**68**	21	11	0
	Médio	61	**27**	10	2	64	**25**	10	1	68	**21**	10	1
	Alto	10	40	**44**	6	13	42	**43**	2	13	42	**40**	5
	Grito	0	0	3	**97**	0	0	5	**95**	0	0	13	**87**
		Taxa de classificação média: **61 ,0**				Taxa de classificação média: **59,5**				Taxa de classificação média: **54,0**			
MFCC	Condição	Cond. Classificada com GMM				Cond. Classificada com HMM				Cond. Classificada com SVM			
	Real	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.
	Neutro	**49**	25	24	2	**50**	24	24	2	**41**	28	27	4
	Médio	35	**26**	37	2	37	**25**	32	6	44	**18**	32	6
	Alto	25	9	**64**	2	23	12	**62**	3	24	13	**59**	4
	Grito	0	0	2	**98**	0	1	2	**97**	0	1	10	**89**
		Taxa de classificação média: **59,3**				Taxa de classificação média: **58,5**				Taxa de classificação média: **51,8**			
TEO	Condição	Cond. Classificada com GMM				Cond. Classificada com HMM				Cond. Classificada com SVM			
	Real	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.
	Neutro	**38**	33	19	10	**37**	33	19	11	**28**	36	22	14
	Médio	29	**35**	31	5	30	**33**	31	6	34	**25**	34	7
	Alto	11	20	**63**	6	12	20	**60**	8	12	20	**56**	12
	Grito	0	0	1	**98**	0	0	4	**96**	0	0	11	**89**
		Taxa de classificação média: **58,5**				Taxa de classificação média: **56,5**				Taxa de classificação média: **49,5**			

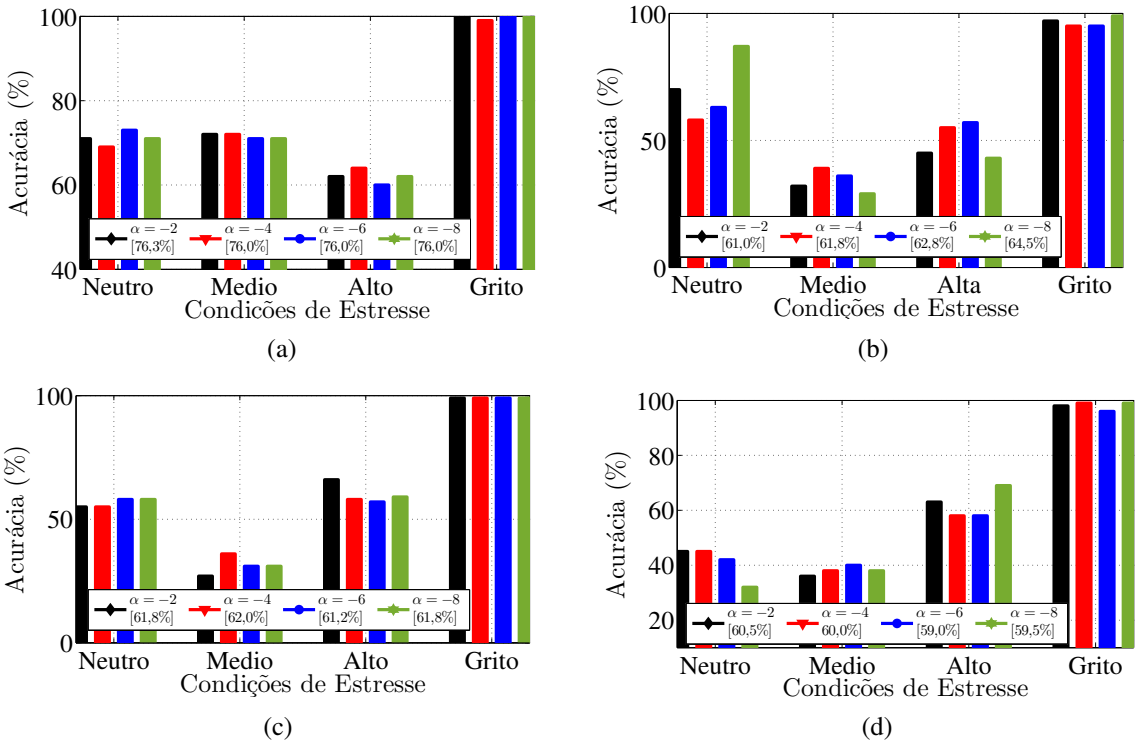


Figura 4.6 – Acurácia obtida da classificação utilizando α-GMM com a base SUSAS, para os seguintes atributos: (a) HHHC; (b) pH; (c) MFCC; (d) TEO.

o classificador α-GMM obteve desempenho superior aos classificadores clássicos. No contexto das diferentes bases utilizadas, o atributo proposto, HHHC. Isto indica o potencial do HHHC

independente da língua e do contexto de gravação dos sinais acústicos. A base EMO-DB obteve os valores mais elevados de acurácia média máxima, o que pode ser devido ao fato de ser uma base com gravações mais controladas do que as demais. Em relação aos valores de α, HHHC tem valores mais elevados de acurácia média entre $-8 \leq \alpha \leq -4$ no contexto da classificação de emoções (bases EMO-DB, IEMOCAP, SEMAINE e RECOLA).

Tabela 4.6 – Resumo dos melhores resultados de classificação.

Base EMO-DB		
Atributo	Acurácia média máxima	Classificador
HHHC	$79,2\%$	α-GMM (-4, -6 e -8)
pH	$65,4\%$	α-GMM (-2)
MFCC	$63,6\%$	α-GMM (-8)
TEO	$52,8\%$	α-GMM (-2 e -4)
Base IEMOCAP		
Atributo	Acurácia média máxima	Classificador
HHHC	$61,0\%$	α-GMM (-8)
pH	$52,8\%$	α-GMM (-8)
MFCC	$50,8\%$	α-GMM (-4)
TEO	$44,2\%$	α-GMM (-6)
Base SEMAINE		
Atributo	Acurácia média máxima	Classificador
HHHC	$54,5\%$	α-GMM (-6)
pH	$50,8\%$	α-GMM (-4)
MFCC	$49,0\%$	α-GMM (-6)
TEO	$40,8\%$	α-GMM (-8)
Base RECOLA		
Atributo	Acurácia média máxima	Classificador
HHHC	$63,0\%$	α-GMM (-6)
pH	$59,5\%$	α-GMM (-4)
MFCC	$54,5\%$	α-GMM (-4)
TEO	$49,5\%$	α-GMM (-2)
Base SUSAS		
Atributo	Acurácia média máxima	Classificador
HHHC	$76,3\%$	α-GMM (-2)
pH	$64,5\%$	α-GMM (-8)
MFCC	$62,0\%$	α-GMM (-4)
TEO	$60,5\%$	α-GMM (-2)

4.3.7 – Resultados HHHC+INS

Os resultados do processo de classificação utilizando o INS como informação adicional ao vetor HHHC são apresentados a seguir. No contexto do α-GMM, é colocado o valor de α que proporcionou o melhor desempenho no caso do atributo HHHC para cada base acústica.

Na Tabela 4.7 estão os resultados obtidos para a fusão HHHC+INS com a base EMO-DB. Em relação ao atributo HHHC individualmente, a acurácia média para HHHC+INS foi maior considerando todos os classificadores utilizados. O aumento em relação ao α-GMM ($\alpha = -6$) foi de 2,6 p.p.. No caso dos outros classificadores, a diferença entre HHHC+INS para HHHC foi de 2,2 p.p. com HMM e 1,6 p.p. com SVM. Em consequência, os resultados HHHC+INS são ainda maiores que os atributos comparativos apresentados na Seção 4.3.1.

Os resultados de HHHC+INS para a base IEMOCAP são mostrados na Tabela 4.8. A melhora na classificação de cada emoção individualmente foi de aproximadamente 2 p.p..

Tabela 4.7 – Taxas de acurácia (%) de 5 estados emocionais para HHHC+INS considerando os classificadores α-GMM, HMM e SVM com a base EMO-DB.

	Emoção	Emoção Classificada com α-GMM					Emoção Classificada com HMM					Emoção Classificada com SVM				
	Real	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.	Rai.	Fel.	Neu.	Téd.	Tri.
HHHC+INS	Raiva	**88**	12	0	0	0	**77**	23	0	0	0	**73**	27	0	0	0
	Felicidade	32	**68**	0	0	0	30	**70**	0	0	0	36	**64**	0	0	0
	Neutro	0	0	**87**	13	0	0	0	**84**	16	0	0	0	**67**	23	0
	Tédio	0	0	10	**77**	13	0	0	14	**71**	15	0	0	19	**52**	29
	Tristeza	0	0	0	11	**89**	0	0	0	18	**82**	0	0	0	27	**73**
		Taxa de classificação média: **81,8**					Taxa de classificação média: **76,8**					Taxa de classificação média: **65,8**				

No caso do classificador α-GMM, a acurácia média foi de $61,0\%$ com HHHC para $62,8\%$ com HHHC+INS, considerando $\alpha = -8$. Para o HMM, ocorre um aumento de até 4 p.p. na identificação da emoção tristeza. Neste contexto, a taxa de acerto média foi de $52,0\%$ com HHHC para $55,3\%$ com HHHC+INS. Na classificação com SVM, a acurácia média com HHHC+INS foi 1,7 p.p. maior que aquela obtida com HHHC.

Tabela 4.8 – Taxas de acurácia (%) de 4 estados emocionais para HHHC+INS considerando os classificadores α-GMM, HMM e SVM com a base IEMOCAP.

	Emoção	Emoção Classificada com α-GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.
HHHC+INS	Raiva	**68**	23	9	0	**58**	28	13	1	**51**	31	14	4
	Felicidade	26	**57**	15	2	30	**48**	18	4	30	**38**	27	5
	Neutro	9	11	**63**	17	11	13	**57**	19	15	19	**40**	26
	Tristeza	6	9	22	**63**	6	10	26	**58**	7	14	32	**47**
		Taxa de classificação média: **62,8**				Taxa de classificação média: **55,3**				Taxa de classificação média: **44,0**			

No contexto da base SEMAINE, seus resultados estão apresentados na Tabela 4.9. A fusão HHHC+INS tem taxas de acerto superiores aos resultados obtidos com HHHC (apresentados na Seção 4.3.3) para todos os classificadores empregados os experimentos. Com α-GMM, o INS agregou um aumento de 2,5 p.p. em termos de acurácia média, considerando um valor de α igual a -6. Em relação aos outros classificadores, a mesma diferença de 2,5 p.p. foi observada no caso do uso de HMM, e, com SVM, o aumento na taxa de acerto com HHHC+INS foi de 1,5 p.p. em relação ao vetor HHHC.

Tabela 4.9 – Taxas de acurácia (%) de 4 estados emocionais para HHHC+INS considerando os classificadores α-GMM, HMM e SVM com a base SEMAINE.

	Emoção	Emoção Classificada com α-GMM				Emoção Classificada com HMM				Emoção Classificada com SVM			
	Real	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.
HHHC + INS	Raiva	**51**	23	20	6	**46**	25	22	7	**41**	28	24	7
	Felicidade	14	**59**	25	2	17	**53**	28	2	19	**45**	31	5
	Diversão	13	24	**55**	8	13	27	**51**	9	15	30	**44**	11
	Tristeza	5	15	17	**63**	5	18	22	**55**	7	20	26	**47**
		Taxa de classificação média: **57,0**				Taxa de classificação média: **51,3**				Taxa de classificação média: **44,3**			

As taxas de acerto obtidas da fusão HHHC+INS no caso da base RECOLA estão na Tabela 4.10. Para o melhor caso de classificação com α-GMM, o INS proporciona ao HHHC um aumento de 1 p.p. na taxa média de acerto média ($\alpha = -6$). Com o HMM, a acurácia média foi de 55% com HHHC para $56,5\%$ com HHHC+INS. A classificação com SVM, por sua vez, teve um aumento de 1,5 p.p. no uso do INS adicionado ao vetor HHHC.

Tabela 4.10 – Taxas de acurácia (%) em relação ao nível de ativação dos estados emocionais para HHHC+INS considerando os classificadores α-GMM, HMM e SVM com a base RECOLA.

	Emoção	Classificação com α-GMM		Classificação com HMM		Classificação com SVM	
HHHC+INS	Real	Alta	Baixa	Alta	Baixa	Alta	Baixa
	Alta ativação	**65**	35	**58**	42	**50**	50
	Baixa ativação	37	**63**	45	**55**	48	**52**
		Taxa de classificação média: **64,0**		Taxa de classificação média: **56,5**		Taxa de classificação média: **51,0**	

Em relação a base SUSAS, os resultados de HHHC+INS são apresentados na Tabela 4.11. O aumento na acurácia média em relação ao α-GMM foi com um valor diferente de α. Com INS, a taxa de acerto chegou a $78,3\%$ considerando $\alpha = -4$, contra $76,3\%$ do vetor HHHC com $\alpha = -2$. No caso do classificador HMM, ocorre um aumento de 2,8 p.p. na acurácia de HHHC+INS em relação a HHHC. Com SVM, esta diferença é de 1,5 p.p.. Note que, assim como ocorreu com os atributos individualmente, no caso do HHHC+INS os classificadores estocásticos obtiveram melhor desempenho em relação ao classificador não paramétrico.

Tabela 4.11 – Taxas de acurácia (%) de 4 condições de estresse para HHHC+INS considerando os classificadores α-GMM, HMM e SVM com a base SUSAS.

	Condição	Cond. Classificada com α-GMM				Cond. Classificada com HMM				Cond. Classificada com SVM			
HHHC + INS	Real	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.	Neu.	Med.	Alt.	Gri.
	Neutro	**70**	21	8	1	**71**	21	8	0	**64**	25	11	0
	Médio	15	**75**	10	0	17	**63**	20	0	22	**54**	24	0
	Alto	2	20	**68**	10	4	20	**64**	12	7	23	**51**	19
	Grito	0	0	0	**100**	0	0	1	**99**	0	0	9	**91**
		Taxa de classificação média: **78,3**				Taxa de classificação média: **74,3**				Taxa de classificação média: **65,0**			

4.3.8 – Fusão de Atributos

Nos experimentos realizados, buscou-se verificar a contribuição do vetor HHHC em relação aos três atributos comparativos. O objetivo disto é investigar um aumento nas taxas de acerto de pH, MFCC e TEO quando submetidos à fusão com o HHHC. Uma vez que o α-GMM foi o classificador com os melhores resultados, ele foi escolhido para esta tarefa.

Na Figura 4.7 são mostrados os resultados obtidos do procedimento de fusão de atributos utilizados na base EMO-DB. No que diz respeito à fusão pH+HHHC, a maior taxa de acurácia média foi obtida pelo classificador α-GMM considerando $\alpha = -6$ $(75,6\%)$. Isto representa um aumento de aproximadamente 10 p.p. em relação ao melhor resultado obtido com o vetor pH ($65,4\%$, com $\alpha = -2$). Note que, mesmo considerando $\alpha = -2$ (Figura 4.7a), pH+HHHC atinge uma taxa de acerto quase 9 p.p. acima daquela obtida com o pH. No contexto da análise individual de cada estado emocional, a maior contribuição do vetor HHHC ao vetor pH foi verificada no caso da emoção Felicidade, em que teve uma melhora de 67% ($\alpha = -4$) com pH para 82% ($\alpha = -8$) com pH+HHHC. A fusão MFCC+HHHC obtém sua maior taxa de acurácia média considerando $\alpha = -8$ $(73,7\%)$. Isto significa que o vetor HHHC proporcionou uma melhora de quase 10 p.p. ao atributo MFCC no contexto da EMO-DB ($63,6\%$, com $\alpha = -8$). Para esta fusão, pode ser observado um aumento de 31 p.p. na classificação do estado Neutro,

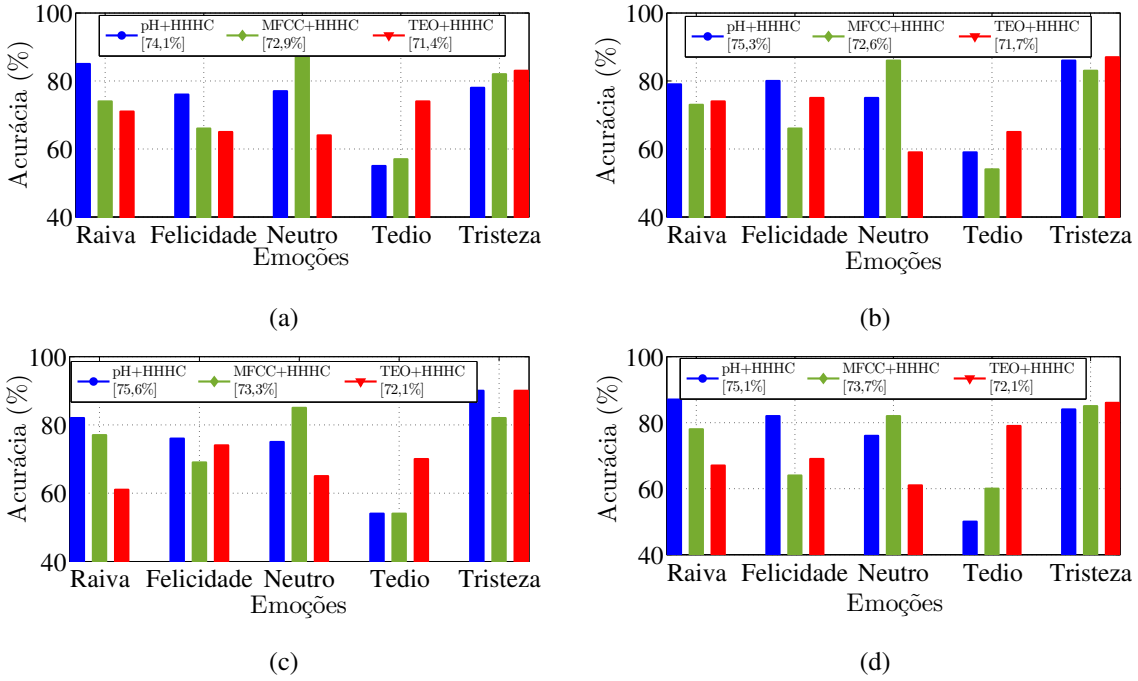


Figura 4.7 – Acurácia obtida da fusão de atributos com a base EMO-DB, utilizando α-GMM: (a) $\alpha = -2$; (b) $\alpha = -4$; (c) $\alpha = -6$; (d) $\alpha = -8$.

ou seja, MFCC+HHHC atinge 86% ($-6 \leq \alpha \leq -2$) contra 55% obtido com MFCC. Enquanto que o estado de Tédio é classificado com 37% de acurácia com MFCC, MFCC+HHHC obtém 56% de acerto com $\alpha = -8$. Outra contribuição significativa de HHHC ao MFCC foi em relação à emoção Tristeza. Para este estado emocional, HHHC proporciona ao atributo MFCC um aumento de 68% a 86% (ambos usando $\alpha = -8$). Em relação à fusão TEO+HHHC, a maior taxa de acurácia média foi $72,1\%$, obtida usando $\alpha = -6$ e $\alpha = -8$. Isto significa uma melhora de 19.3 p.p. proporcionada pelo vetor HHHC ao atributo TEO. No caso do estado Neutro, sua taxa de acerto com o TEO foi de 47% ($\alpha = -4$) para 65% ($\alpha = -6$).

Para a base IEMOCAP, os resultados obtidos do processo de fusão entre HHHC e os atributos comparativos são apresentados na Figura 4.8. Na fusão pH+HHHC, é atingido uma acurácia média de $63,2\%$ ($\alpha = -8$), o que supera os valores obtidos com pH ($52,8\%$) e HHHC+INS ($62,8\%$). Isto pode ser um indicativo de que a fusão de atributos baseados no expoente de Hurst (pH+HHHC), com sua relação com a PSD, proporciona um alto desempenho na separação de emoções. A maior contribuição do HHHC ao pH é na classificação da emoção Felicidade, a qual aumenta de $47,0\%$ ($\alpha = -6$ e $\alpha = -8$) para $58,0\%$ ($\alpha = -8$). Na fusão MFCC+HHHC, HHHC proporciona ao MFCC um aumento na acurácia média de $50,8\%$ para $60,5\%$ (ambos com $\alpha = -4$). Nesta fusão, a melhora mais significativa diz respeito ao estado Neutro, que teve sua taxa de acerto aumentada de $47,0\%$ ($\alpha = -4$) com MFCC para $60,0\%$ com MFCC+HHHC considerando o mesmo valor de α. Em relação à fusão, a maior taxa média de acerto foi $56,1\%$, obtida com $\alpha = -4$, o que é 11.9 p.p. acima do que foi obtido com o

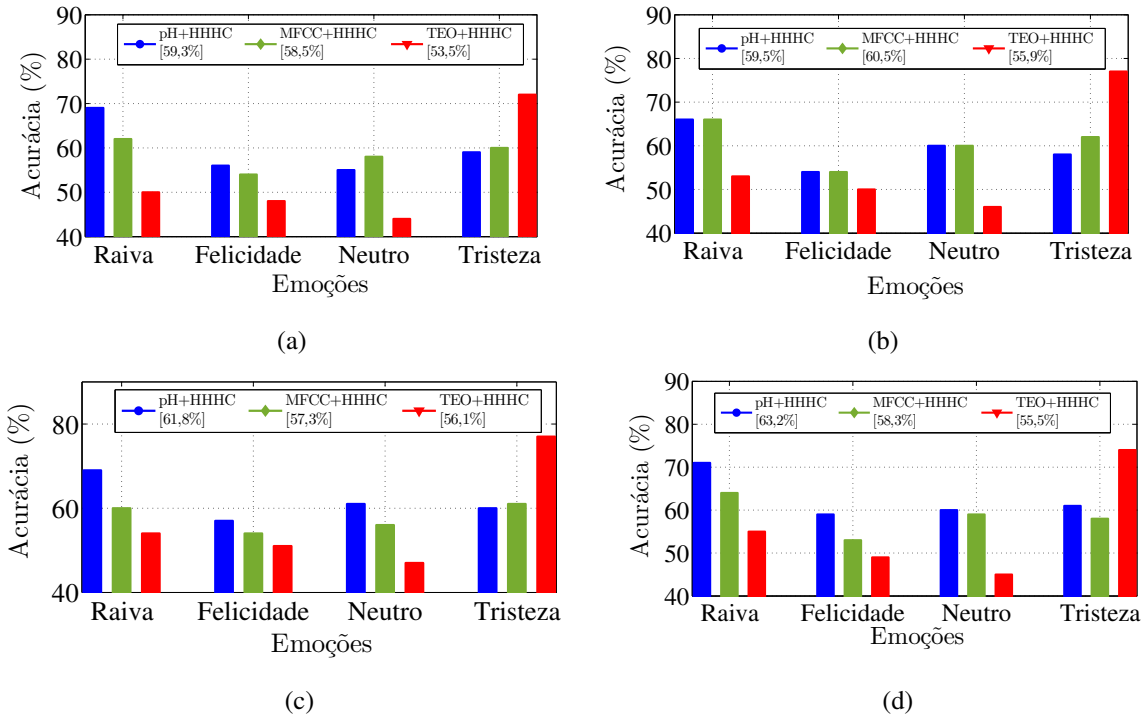


Figura 4.8 – Acurácia obtida da fusão de atributos com a base IEMOCAP, utilizando α-GMM: (a) $\alpha = -2$; (b) $\alpha = -4$; (c) $\alpha = -6$; (d) $\alpha = -8$.

atributo TEO $(44,2\%)$. Para a emoção Raiva houve a maior contribuição do HHHC para TEO, com um aumento de $40,0\%$ para $55,0\%$ na taxa de acerto.

No contexto da base SEMAINE, os resultados da fusão de atributos são apresentados na Figura 4.9. Em relação a fusão pH+HHHC, a maior acurácia média atingida é de $56,5\%$ ($\alpha = -4$), o que representa uma melhora nas taxas de acerto sobre o vetor pH, o vetor HHHC e a combinação HHHC+INS. Individualmente, a maior contribuição do HHHC ao vetor pH é no caso da emoção Diversão, em que pH+HHHC supera pH em 9 p.p. considerando, para ambos os casos, $\alpha = -4$ (de $48,0\%$ para $57,0\%$). Na tarefa de fusão entre MFCC e HHHC (MFCC+HHHC), pode ser observado um aumento na taxa acurácia média da classificação de $49,0\%$ para $53,6\%$, com $\alpha = -6$. Nesta tarefa de fusão, o maior aumento foi em relação à emoção Felicidade, em que MFCC+HHHC atinge uma acurácia média de $58,0\%$ contra $52,0\%$ obtido pelo MFCC. A fusão TEO+HHHC obtém uma taxa de acurácia média de $47,4\%$, o que promove um aumento de 6,6 p.p. em relação ao que foi obtido com o atributo TEO para o mesmo valor de $\alpha = -8$. A maior contribuição do HHHC ao TEO no contexto da SEMAINE foi em relação à emoção Felicidade, em que sua taxa subiu de $33,0\%$ ($\alpha = -8$) para $58,0\%$ ($\alpha = -6$).

Na Figura 4.10 são mostrados os resultados obtidos da classificação com a fusão de atributos extraídos da base RECOLA. No que diz respeito à fusão pH+HHHC, o maior valor de acurácia média ocorre para $\alpha = -4$ $(68,5\%)$, o que representa um aumento de 9 p.p. em relação ao que foi obtido com pH, também para $\alpha = -4$. Na fusão MFCC+HHHC, a maior taxa média

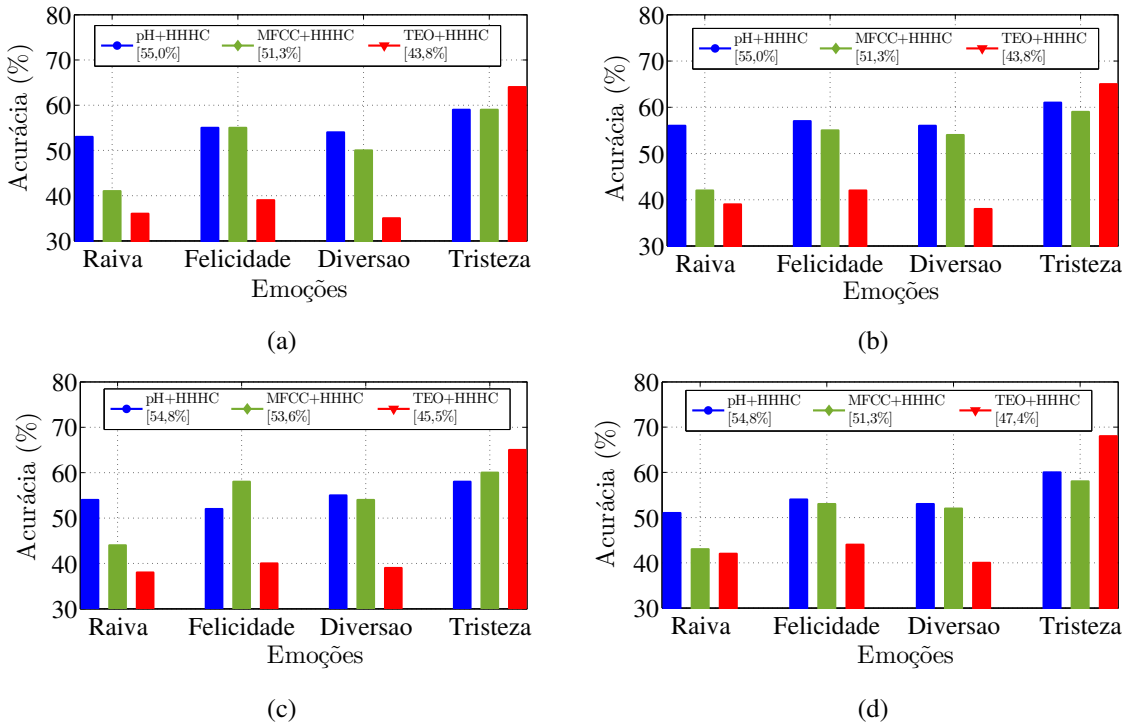


Figura 4.9 – Acurácia obtida da fusão de atributos com a base SEMAINE, utilizando α-GMM: (a) $\alpha = -2$; (b) $\alpha = -4$; (c) $\alpha = -6$; (d) $\alpha = -8$.

de acerto foi obtida utilizando $\alpha = -4$ no classificador α-GMM $(61,5\%)$. Isto significa um aumento de 7 p.p. em relação ao que foi obtido com MFCC $(54,5\%)$, para este mesmo valor de α. No contexto da fusão TEO+HHHC, foi atingida uma acurácia média de $56,0\%$ considerando $\alpha = -2$. Este resultado supera em 6,5 p.p. o que foi obtido com o atributo TEO para este mesmo valor de $\alpha = -2$ $(49,5\%)$.

A fusão de atributos para a base SUSAS tem seus resultados apresentados na Figura 4.11. A fusão pH+HHHC atinge sua maior acurácia média $(75,1\%)$ com $\alpha = -2$. Este resultado supera o desempenho do melhor resultado obtido com pH $(64,5\%$ com $\alpha = -8)$. No caso da classificação da condição de Alto estresse, há um aumento de $43,0\%$ com pH para $80,0\%$ com pH+HHHC. Na fusão MFCC+HHHC, o maior valor de acurácia média $(70,6\%)$ é obtido com $\alpha = -6$. Para a fusão TEO+HHHC, há um aumento de $60,5\%$ $(\alpha = -2)$ para $71,1\%$ $(\alpha = -6)$. O vetor HHHC contribui para um aumento na taxa de acerto de todas as condições de estresse consideradas nos experimentos. O aumento mais significativo foi na classificação de Médio estresse, para o qual TEO+HHHC chega a uma acurácia de $63,0\%$ com $\alpha = -6$. Isto significa 23 p.p. acima da taxa de acerto obtida pelo atributo TEO $(40,0\%)$.

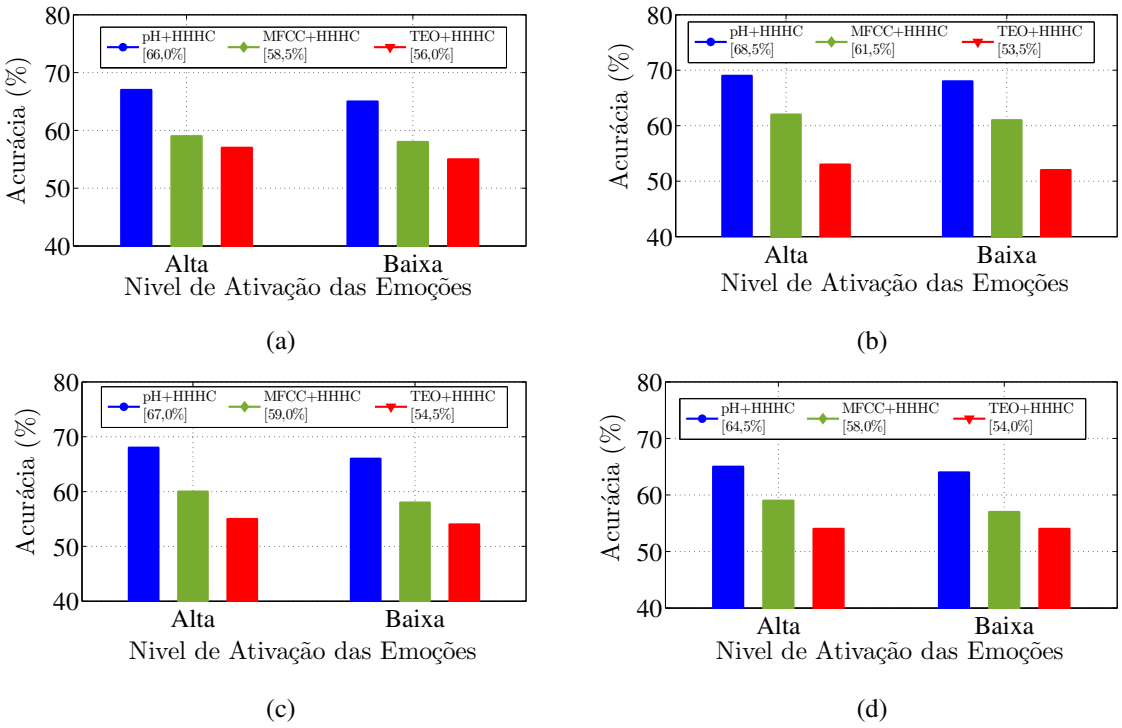


Figura 4.10 – Acurácia obtida da fusão de atributos com a base RECOLA, utilizando α-GMM: (a) $\alpha = -2$; (b) $\alpha = -4$; (c) $\alpha = -6$; (d) $\alpha = -8$.

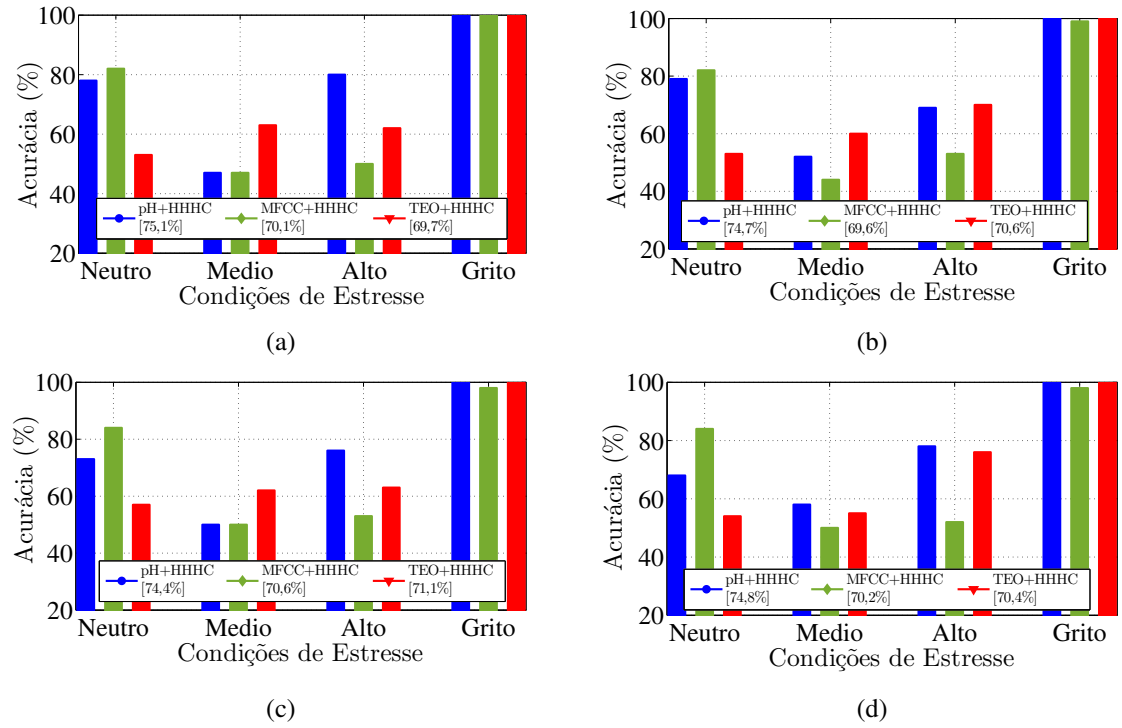


Figura 4.11 – Acurácia obtida da fusão de atributos com a base SUSAS, utilizando α-GMM: (a) $\alpha = -2$; (b) $\alpha = -4$; (c) $\alpha = -6$; (d) $\alpha = -8$.

4.4 – Resumo

Este Capítulo apresentou os experimentos realizados em relação à classificação das variações acústicas afetivas não estacionárias. Para tanto, foram utilizados os seguintes classificadores clássicos: GMM, HMM e SVM. Ainda, foi proposto o classificador α-GMM para a tarefa de classificação de estados emocionais e condições de estresse. Para a avaliação da robustez do atributo acústico HHHC, foram analisadas cinco bases acústicas: EMO-DB, IEMOCAP, SEMAINE, RECOLA e SUSAS. Para fins comparativos, foram utilizados os atributos pH, MFCC E TEO. Os resultados demonstraram que o vetor HHHC atinge desempenho superior em relação aos atributos comparativos. Além disso, a informação do INS agregada ao vetor HHHC (HHHC+INS) proporciona um aumento nas taxas de acerto. Em relação aos classificadores utilizados, foi verificado que o α-GMM obtém os resultados mais significativos. Em relação aos classificadores clássicos, os estocásticos GMM e HMM obtiveram performance superior ao SVM. Isto indica que, para tarefas de classificação com matrizes de atributos de baixa dimensão, os classificadores estocásticos são mais robustos (principalmente o α-GMM). No contexto da fusão de atributos, foi observado que o HHHC agrega valor às taxas de acerto de todos os atributos comparativos. Na análise da fusão considerando o classificador α-GMM, o vetor HHHC proporciona melhora nos valores de acurácia de pH, MFCC e TEO em todas as bases acústicas consideradas nos experimentos.

CAPÍTULO 5

Conclusão e Trabalhos Futuros

Nesta Tese, foram analisadas variações acústicas geradas em diferentes estados emocionais e condições de estresse. Os estados afetivos refletem experiências subjetivas em curtos períodos de tempo que podem ser observadas na fisiologia humana, tais como fisionomia, batimentos cardíacos e fala. A abordagem utilizada levou em consideração a não estacionariedade dessas variações afetivas, que são introduzidas na voz em seu processo de produção.

Para a análise das variações acústicas afetivas não estacionárias, foram utilizados a técnica de decomposição EMD e um índice (INS). O método EMD é apropriado para a decomposição tempo-frequência de sinais não estacionários. Nesta análise, os sinais acústicos foram decompostos em seis faixas de frequência (IMFs) de modo que fosse observado o comportamento das diferentes variações afetivas. O método EEMD foi utilizado como alternativa ao método EMD na análise dos sinais acústicos. Nesta pesquisa, foi observado que a decomposição baseada em EEMD consegue separar mais apropriadamente as faixas de frequência do que a EMD. O INS foi empregado para analisar o grau de não estacionariedade das variações acústicas. Os resultados desta análise demonstraram que os estados afetivos apresentam, em seus sinais acústicos, diferentes graus de não estacionariedade. Além disso, foi observada uma relação entre o INS e a PSD das variações emocionais.

A partir desta análise, foi observado que o comportamento não estacionário das variações afetivas podem ser representadas de acordo com seu específico índice ou grau de não estacionariedade. Isto quer dizer que os estados afetivos possuem impressões biométricas que são observadas por meio de seus sinais acústicos. Para extrair essa informação biométrica dos sinais acústicos, foi utilizado o expoente de Hurst, o qual foi estimado de cada função resultante da decomposição do sinal em análise. Com isso, foi definido um novo vetor de atributos acústico, o HHHC, que se baseia na captura da informação não linear de cada componente das variações acústicas não estacionárias. Os resultados com o vetor HHHC demonstraram a acurácia na classificação dos diferentes tipos de variações afetivas. Em comparação com outros atributos encontrados na literatura, tais como o MFCC, os resultados com HHHC foram superiores. Como informação adicional, foi utilizado o INS estimado de cada IMF, formando

assim a combinação HHHC+INS. Os resultados desta fusão demonstraram que o INS agrega informação útil no vetor HHHC, de modo a aprimorar as taxas de acerto na classificação. Ainda, foi realizada uma etapa de fusão do HHHC com os demais atributos considerados nos experimentos. Os resultados deste procedimento demonstraram que o HHHC aprimora a classificação destes atributos.

Na etapa de classificação das variações acústicas afetivas, foi proposto o α-GMM. Este método foi empregado em princípio na tarefa de reconhecimento de locutor. Nesta Tese, o α-GMM foi utilizado na classificação de emoções e de condições de estresse. Para fins comparativos, foram empregados o clássico GMM, o HMM e o SVM. Os resultados demonstraram que os classificadores baseados em uma abordagem estocástica (α-GMM, GMM e HMM) obtiveram os melhores desempenhos. Ainda, foi observado que o α-GMM obtém desempenho superior aos demais classificadores na tarefa de classificação dos estados afetivos considerados nesta pesquisa.

As principais contribuições deste trabalho de Tese foram as seguintes:

- Análise de variações afetivas com base na não estacionariedade de seus sinais acústicos;
- Abordagem com decomposição tempo-frequência baseada em EMD e utilização do INS. Isto demonstrou que cada variação afetiva possui um grau de não estacionariedade diferente. Além disso, a decomposição adaptativa mostrou que o aspecto da não estacionariedade pode ser uma impressão biométrica de estados afetivos;
- Proposta do atributo acústico HHHC, bem como a fusão HHHC+INS, mostrando robustez na classificação das variações acústicas afetivas e agregando valor a atributos clássicos nas taxas de acerto;
- Proposta do α-GMM para a classificação de variações acústicas afetivas;
- Utilização de bases acústicas de três idiomas diferentes: inglês, alemão e francês. Além disso, o contexto de gravação das bases são diferentes, a exemplo da base EMO-DB (emoções atuadas) e da base SEMAINE (emoções induzidas). Ainda, esta Tese considerou dois tipos diferentes de estados afetivos: o contexto de variações emocionais (bases EMO-DB, IEMOCAP, SEMAINE e RECOLA) e de diferentes condições de estresse (base SUSAS).

5.1 – Sugestão para Trabalhos Futuros

Como sugestões para trabalhos futuros, pode-se destacar as seguintes:

- Empregar ruídos acústicos a fim de investigar a robustez do vetor HHHC na detecção de variações afetivas em condições ruidosas;

- Utilizar métodos que aprimorem os resultados obtidos com o HHHC, tais como máscaras acústicas [7];
- Investigar a aplicação do vetor HHHC na tarefa de verificação de emoções, bem como a fusão HHHC+INS;
- Realizar a fusão do HHHC com outros atributos, tais como o eGeMAPS [31], a fim de investigar o aprimoramento na taxa de acerto proporcionada por eles;
- Investigar a análise das variações acústicas afetivas não estacionárias quando realizada a separação de gêneros. Além disso, empregar o vetor HHHC em outras bases de dados, em outros idiomas;
- Avaliar o impacto das variações acústicas na inteligibilidade;
- Empregar outros classificadores para investigar o desempenho do atributo HHHC, tais como redes neurais profundas (*Deep Neural Networks* – DNN) e redes neurais convolutivas (*Convolutional Neural Networks* – CNN).

5.2 – Comentários Finais

Nesta Tese foram analisadas variações acústicas provocadas por diferentes estados afetivos. Neste contexto, foram utilizadas as técnicas EMD e INS, com as quais foi observado que a informação baseada na não estacionariedade dessas variações acústicas por ser uma informação biométrica. Então, foi proposto o atributo HHHC, que realiza uma decomposição baseada em EMD e extrai informação de cada IMF utilizando o expoente de Hurst. O INS foi utilizado como informação adicional, provendo uma melhora nas taxas de acerto do vetor HHHC. Na tarefa de classificação, foi proposto o α-GMM, com o qual foram obtidas taxas de acurácia superiores aos classificadores clássicos encontrados na literatura. Dessa forma, os resultados obtidos dos experimentos realizados nesta pesquisa demonstraram que o atributo proposto é capaz de extrair informação característica de cada estado afetivo, sendo assim robusto no processo de classificação.

Referências Bibliográficas

[1] R. Plutchik, *The emotions*. University Press of America, 1991.

[2] B. Yang and M. Lugger, "Emotion Recognition from Speech Signals using New Harmony Features," *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, 2010.

[3] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. IEEE New York, NY, USA:, 2000.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[6] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *IEEE Workshop on Automatic Speech Recognition & Understanding, 2009. ASRU 2009*, pp. 552–557, IEEE, 2009.

[7] L. Zão, D. Cavalcante, and R. Coelho, "Time-Frequency Feature and AMS-GMM Mask for Acoustic Emotion Classification," *IEEE Signal Processing Letters*, vol. 21, pp. 620–624, May 2014.

[8] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Exploitation of Phase-Based Features for Whispered Speech Emotion Recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2016.

[9] M. Tahon and L. Devillers, "Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 1, pp. 16–28, 2016.

[10] K. R. Scherer, "Vocal Communication of Emotion: A Review of Research Paradigms," *Speech Communication*, vol. 40, no. 1, pp. 227–256, 2003.

[11] P. Ekman, *The Handbook of Cognition and Emotion*, ch. Basic Emotions, pp. 45–60. Wiley Online Library, 1999.

[12] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear Feature based Classification of Speech Under Stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[13] J. M. Pickett, "The Sounds of Speech Communication," *Baltimore, MD: University Park*, 1980.

[14] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education, 2002.

[15] N. Wang, P. Ching, N. Zheng, and T. Lee, "Robust Speaker Recognition using Denoised Vocal Source and Vocal Tract Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 196–205, 2011.

[16] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, March 1998.

[17] H. E. Hurst, "Long-Term Storage Capacity of Reservoirs," *Trans. Amer. Soc. Civil Eng.*, vol. 116, pp. 770–808, 1951.

[18] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing Stationarity with Surrogates: A Time-Frequency Approach," *IEEE Transactions on Signal Processing*, vol. 58, pp. 3459–3470, July 2010.

[19] D. Wu, J. Li, and H. Wu, "α-Gaussian Mixture Modelling for Speaker Recognition," *Pattern Recognition Letters*, vol. 30, no. 6, pp. 589–594, 2009.

[20] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[21] L. Rabiner and B. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[22] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[23] C. Busso, S. Lee, and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, 2009.

[24] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion Recognition using a Hierarchical Binary Decision Tree Approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[25] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, and D. D. O'Shaughnessy, "Amplitude Modulation Features for Emotion Recognition from speech," in *Proc. INTERSPEECH, 2013*, pp. 2420–2424, 2013.

[26] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and Emotion Classification using Jitter and Shimmer Features," in *Proc. ICASSP, 2007*, vol. 4, pp. IV–1081, IEEE, 2007.

[27] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[28] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech Emotion Recognition using Fourier Parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[29] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic Feature Selection for Automatic Emotion Recognition from Speech," *Information Processing & Management*, vol. 45, no. 3, pp. 315–328, 2009.

[30] W. Zheng, M. Xin, X. Wang, and B. Wang, "A Novel Speech Emotion Recognition Method via Incomplete Sparse Least Square Regression," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014.

[31] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, and S. S. Narayanan, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[32] H. Teager and S. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract," in *Speech Production and Speech Modelling*, pp. 241–261, Springer, 1990.

[33] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," International Conference on Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990.

[34] K. R. Scherer, "Psychological Models of Emotion," *The Neuropsychology of Emotion*, vol. 137, no. 3, pp. 137–162, 2000.

[35] S. Steidl, *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*. University of Erlangen-Nuremberg Germany, 2009.

[36] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," in *Proc. EUROSPEECH, 1997*, no. 4, pp. 1743–46, 1997.

[37] D. Konstan, "A Raiva e as Emoções em Aristóteles: as Estratégias do Status," *Letras Clássicas*, no. 4, pp. 77–90, 2000.

[38] A. Damásio, *O Erro de Descartes: Emoção, Razão e o Cérebro Humano*. Editora Companhia das Letras, 2012.

[39] C. Darwin and P. Prodger, *The Expression of the Emotions in Man and Animals*. Oxford University Press, USA, 1998.

[40] R. Plutchik, "A General Psychoevolutionary Theory of Emotion," *Theories of emotion*, vol. 1, no. 3-31, p. 4, 1980.

[41] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[42] H. Schlosberg, "Three Dimensions of Emotion," *Psychological Review*, vol. 61, no. 2, p. 81, 1954.

[43] W. James, "What is an emotion?," *Mind*, vol. 9, no. 34, pp. 188–205, 1884.

[44] J. Cai, G. Liu, and M. Hao, "The Research on Emotion Recognition from ECG Signal," in *International Conference on Information Technology and Computer Science, 2009. ITCS 2009.*, vol. 1, pp. 497–500, IEEE, 2009.

[45] M. Murugappan, N. Ramachandran, and Y. Sazali, "Classification of Human Emotion from EEG using Discrete Wavelet Transform," *Journal of Biomedical Science and Engineering*, vol. 3, no. 04, p. 390, 2010.

[46] A. Kumar and S. K. Mullick, "Nonlinear Dynamical Analysis of Speech," *The Journal of the Acoustical Society of America*, vol. 100, p. 615, 1996.

[47] W. Martin, "Measuring the Degree of Non-Stationarity by using the Wigner-Ville Spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'84.*, vol. 9, pp. 262–265, IEEE, 1984.

[48] H. Laurent and C. Doncarli, "Stationarity Index for Abrupt Changes Detection in the Time-Frequency Plane," *IEEE Signal Processing Letters*, vol. 5, no. 2, pp. 43–45, 1998.

[49] S. Kay, "A New Nonstationarity Detector," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1440–1451, 2008.

[50] P. Basu, D. Rudoy, and P. J. Wolfe, "A Nonparametric Test for Stationarity based on Local Fourier Analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, pp. 3005–3008, IEEE, 2009.

[51] M. Basseville, "Distance Measures for Signal Processing and Pattern Recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, 1989.

[52] R. Coelho and L. Zão, "Empirical Mode Decomposition Theory Applied to Speech Enhancement," in *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering and Processing* (R. Coelho, V. Nascimento, R. Queiroz, J. Romano, and C. Cavalcante, eds.), pp. 123–153, Boca Raton, FL: CRC Press, 2015.

[53] P. Flandrin, G. Rilling, and P. Gonçalvès, "Empirical Mode Decomposition as a Filter Bank," *IEEE Signal Processing Letters*, vol. 11, pp. 112–114, February 2004.

[54] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. INTERSPEECH, 2005*, pp. 1517–1520, 2005.

[55] D. P. Mandic, N. ur Rehman, Z. Wu, and N. E. Huang, "Empirical Mode Decomposition-based Time-Frequency Analysis of Multivariate Signals: the Power of Adaptive Data Analysis," *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 74–86, 2013.

[56] Z. Wu and N. E. Huang, "Ensemble Empirical Mode Decomposition: a Noise-Assisted Data Analysis Method," *Advances in Adaptive Data Analysis*, vol. 1, no. 01, pp. 1–41, 2009.

[57] N. Pustelnik, P. Borgnat, and P. Flandrin, "Empirical Mode Decomposition Revisited by Multicomponent Non-Smooth Convex Optimization," *Signal Processing*, vol. 102, pp. 313–331, 2014.

[58] Z. Wu and N. E. Huang, "A Study of the Characteristics of White Noise using the Empirical Mode Decomposition Method," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 460, pp. 1597–1611, The Royal Society, 2004.

[59] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4144–4147, IEEE, 2011.

[60] N. Rehman and D. P. Mandic, "Multivariate Empirical Mode Decomposition," in *Proceedings of The Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 466, pp. 1291–1302, The Royal Society, 2010.

[61] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features, and Methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[62] M. Sondhi, "New Methods of Pitch Extraction," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 262–266, 1968.

[63] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average Magnitude Difference Function Pitch Extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[64] R. Sant'Ana, R. Coelho, and A. Alcaim, "Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 931–940, 2006.

[65] I. Daubechies, *Ten lectures on wavelets*, vol. 61. Society for Industrial and Applied Mathematics, 1992.

[66] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 878–897, 1999.

[67] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[68] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, vol. 100. Prentice-hall Englewood Cliffs, 1978.

[69] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech Emotion Recognition using Hidden Markov Models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[70] S. Sharma and P. Singh, "Emotion Recognition based on Audio Signal using GFCC Extraction and BPNN Classification," *International Journal of Computational Engineering Research*, pp. 39–42.

[71] S. Mohanty, "Language Independent Emotion Recognition in Speech Signals," *International Journal*, vol. 6, no. 10, 2016.

[72] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[73] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[74] D. O'shaughnessy, *Speech communication: human and machine*. Universities press, 1987.

[75] L. A. F. Mendoza, "Redes neurais e máquinas de vetor de suporte no reconhecimento de locutor usando coeficientes mfc e características do sinal glotal," *Universidade Federal Fluminense. Dissertação de Mestrado, 129 p.*, 2009.

[76] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.*, vol. 4, pp. IV–277, IEEE, 2007.

[77] R. D. Patterson, J. Holdsworth, and M. Allerhand, "Auditory models as preprocessors for speech recognition," *The Auditory Processing of Speech: from Auditory Periphery to Words*, pp. 67–89, 1992.

[78] R. Schluter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.*, vol. 4, pp. IV–649, IEEE, 2007.

[79] S. E. Bou-Ghazale and J. H. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.

[80] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Formant position based weighted spectral features for emotion recognition," *Speech Communication*, vol. 53, no. 9, pp. 1186–1197, 2011.

[81] G. Trogeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolucional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204, 2016.

[82] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum Autoencoder-Based Domain Adaptation for Speech Emotion Recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.

[83] L. He, M. Lech, N. C. Maddage, and N. B. Allen, "Study of Empirical Mode Decomposition and Spectral Analysis for Stress and Emotion Classification in Natural Speech," *Biomedical Signal Processing and Control*, vol. 6, no. 2, pp. 139–146, 2011.

[84] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967.

[85] P. Sukhummek, S. Kasuriya, T. Theeramunkong, C. Wutiwiwatchai, and H. Kunieda, "Feature selection experiments on emotional speech classification," in *12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015*, pp. 1–4, IEEE, 2015.

[86] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[87] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," in *Inequalities III: Proceedings of the 3rd Symposium on Inequalities* (O. Shisha, ed.), pp. 1–8, University of California, Los Angeles: Academic Press, 1972.

[88] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[89] A. Milton, S. S. Roy, and S. T. Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[90] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: a stepwise procedure for building and training a neural network," in *Neurocomputing*, pp. 41–50, Springer, 1990.

[91] U. Kressel, "Pairwise classification and support vector machines," *Advances in kernel methods: support vector learning*, pp. 255–268, 1998.

[92] A. Venturini, L. Zao, and R. Coelho, "On Speech Features Fusion, α-Integration Gaussian Modeling and Multi-Style Training for Noise Robust Speaker Classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1951–1964, 2014.

[93] L. Zao and R. Coelho, "Noise Robust Speaker Verification based on the MFCC and pH Features Fusion and Multicondition Training.," in *BIOSIGNALS*, pp. 137–143, 2012.

[94] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, and D. Povey, "The HTK book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[95] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[96] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013*, pp. 1–8, IEEE, 2013.

[97] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[98] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[99] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," in *Eurospeech*, vol. 97, pp. 1743–46, 1997.