Classificação de Variações Acústicas Emocionais com Atributos da Fonte e do Trato Vocal

V. Vieira, R. Coelho e F. M. de Assis

***Resumo*— Este artigo apresenta um estudo sobre a classificação de múltiplas variações acústicas emocionais utilizando os seguintes atributos: vetor de coeficientes de Hurst (pH), coeficientes MFCC (*Mel-Frequency Cepstral Coefficients*) e coeficientes GFCC (*Gammatone-Frequency Cepstral Coefficients*). Para as análises, são empregadas duas bases de dados no idioma inglês, gravadas em diferentes contextos. A classificação é realizada por meio de dois classificadores: GMM (*Gaussian Mixture Models*) e HMM (*Hidden Markov Models*). Os resultados indicam que o atributo da fonte de excitação (pH) é mais eficiente do que aqueles do trato vocal (MFCC e GFCC) em caracterizar as variações emocionais. Em relação aos classificadores, o GMM foi o mais eficiente em modelar cada um dos estados emocionais.**

***Palavras-Chave*— Classificação de emoções, Atributos acústicos, Vetor pH, MFCC, GMM.**

***Abstract*— This article presents a study on the classification of multiple emotional acoustic variations using the following features: vector of Hurst coefficients (pH), Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone-Frequency Cepstral Coefficients (GFCC). For the analysis, two databases are used in English language, recorded in different contexts. The classification is performed by employing two classifiers: Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). Results indicate that the excitation source feature (pH) is more efficient than the vocal tract ones (MFCC and GFCC) in characterizing the emotional variations. Regarding the classifiers, the GMM was the most efficient in modeling each emotional state.**

***Keywords*— Emotion classification, Acoustic features, pH vector, MFCC, GMM.**

I. Introdução

Variações acústicas emocionais estão presentes no cotidiano de comunicação dos seres humanos, influenciando também outros aspectos como a cognição, a percepção e o aprendizado. A classificação dessas variações tem recebido atenção nos últimos anos [1], [2], [3], [4], principalmente por conta de aplicações pertinentes, tais como sistemas de segurança, tradução automática e interação homem-robô [5], [6]. Nas iterações sociais, há uma grande variedade de estados emocionais (a exemplo de Raiva, Felicidade e Tristeza) [7]. De acordo com Ekman [8], há certas emoções que são naturalmente reconhecidas pelos seres humanos. Embora haja essa universalidade na discriminação de estados emocionais, a sua representação em um sistema homem-máquina ainda é um grande desafio [1], [6].

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 [4], [9], [10]. A análise e classificação de variações acústicas emocionais para a definição de um sistema que extraia uma "impressão acústica emocional" 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 [1]; e o ponto de vista dos efeitos das variações afetivas (fonte de excitação glotal *versus* trato vocal, por exemplo).

Alguns estudos utilizam atributos do trato vocal na classificação de emoções, tais como MFCC (*Mel-Frequency Cepstral Coefficients*) e GFCC (*Gammatone-Frequency Cepstral Coefficients*) [3], [11], [12], enquanto outros abordam a proposta de atributos da fonte de excitação glotal, tais como Frequência Fundamental (F0) e o vetor de coeficientes de Hurst (pH) [1], [4]. Em [4], foi observado que o vetor pH é um atributo da fonte vocal que tem relação com a densidade espectral de potência da voz sob efeito de emoções. Os autores empregaram o vetor pH em duas bases de dados: uma de emoções atuadas e outra com diferentes condições de estresse. Em comparação com os coeficientes MFCC, o vetor pH demonstrou desempenho superior nos dois contextos de análise.

Neste trabalho, o vetor pH e os coeficientes MFCC são empregados em duas bases acústicas diferentes daquelas apresentadas em [4]. Em vez de sentenças atuadas, as bases consideradas neste estudo referem-se a diferentes contextos de fala no que diz respeito a espontaneidade e indução das emoções. Adicionalmente, é analisado o desempenho do GFCC como atributo do trato vocal, uma vez que este atributo tem demonstrado desempenho superior ao MFCC em ambientes de variações acústicas, tais como ruídos [13]. A classificação é realizada utilizando modelos de misturas Gaussianas (*Gaussian Mixture Models* – GMM) e modelos de Markov escondidos (*Hidden Markov Models* – HMM). Os resultados dos experimentos mostram que o vetor pH supera os atributos do trato vocal, o que indica que atributos da fonte de excitação glotal são mais adequados para classificar emoções em diferentes contextos de fala. Ainda, a classificação multiestilo mostra que o GMM tem o maior potencial para modelar cada estado emocional individualmente.

O restante deste artigo está organizado da seguinte forma: Na Seção II são apresentadas os atributos acústicos da fonte

V. Vieira, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Campina Grande (PPGEE/UFCG) e atualmente em Pós-Doutorado no Programa de Pós-Graduação em Linguística (PROLING), Universidade Federal da Paraíba (UFPB); R. Coelho, Laboratório de Processamento de Sinais Acústicos, Instituto Militar de Engenharia (IME); F. M. de Assis, Departamento de Engenharia Elétrica, UFCG, Brasil. E-mails: {viniciusjdv@gmail.com, coelho@ime.eb.br, fmarcos@dee.ufcg.edu.br}. Este trabalho foi parcialmente financiado pelo CNPq (140816/2014-3 e 307866/2015-7).

e do trato vocal utilizados neste trabalho. Na Seção III são apresentados os materiais e métodos utilizados neste estudo, no que diz respeito às bases de dados, extração dos atributos acústicos e classificadores. Na Seção IV são apresentados os resultados obtidos nos experimentos realizados, e na Seção V é apresentada a conclusão.

II. Atributos Acústicos

Os atributos acústicos empregados neste trabalho são apresentados nesta Seção. Neste contexto, é considerado um atributo da fonte de excitação glotal (vetor pH) e dois atributos do trato vocal (MFCC e GFCC). Medidas que descrevem os efeitos de variações acústicas na fonte de excitação glotal são comumente chamadas de atributos da fonte de excitação ou atributos da fonte vocal. Atributos do trato vocal capturam características do sinal de voz no domínio da frequência. Por isso, são também conhecidos como atributos espectrais [1], [14]. 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 [15].

A. Vetor pH

O vetor de coeficientes de Hurst (pH) foi proposto inicialmente em [16] como atributo para reconhecimento de locutor. Posteriormente, em [4] o vetor pH foi empregado 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*) [16], 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 [17] são utilizados na DWT;
- Estimação do expoente de Hurst (EH) [18]: 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.

B. MFCC

Os atributos MFCC foram propostos e largamente utilizados em tarefas de reconhecimento de fala e de locutor [19], [20], [21], [22] 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 [23], [24]. Como referência, 1 kHz equivale a 1.000 mels. A transformação de uma frequência f para a escala mel é descrita na Equação 1 [24].

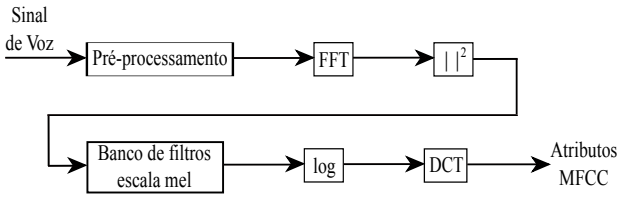

Fig. 1. Extração dos coeficientes MFCC.

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

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 [23], [24].

A extração de atributos MFCC está ilustrada na Figura 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 [19], [20]:

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

em que $j = 1, 2, ..., D$, 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.

C. GFCC

Os atributos GFCC foram propostos para tarefas de reconhecimento de locutor [25], e recentemente têm sido utilizados em reconhecimento de emoções [12]. 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 [26].

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 [27]

$$g(t) = Kt^{(n-1)}e^{-2\pi Bt}\cos(2\pi f_c + \varphi), \quad t > 0, \quad (3)$$

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):

$$G_m = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} (\log Y_n) \cos\left[\frac{\pi m}{N}\left(n - \frac{1}{2}\right)\right], \quad (4)$$

em que $1 \leq m \leq M$, 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.

III. Cenário Experimental

Nesta seção, são apresentados os materiais e métodos empregados neste trabalho. Múltiplos estados emocionais são analisados de diferentes bases de dados, em diferentes contextos de gravação dos sinais acústicos. Para a classificação, são utilizados modelos estocásticos GMM e HMM.

A. Bases de Dados

As bases de dados utilizadas são: *Interactive Realal Dyadic Motion Capture* (IEMOCAP) [28] e *Sustained Realally colored Machine-human Interaction using Nonverbal Expression* (SEMAINE) [29], ambas descritas a seguir.

1) IEMOCAP: Esta base consiste de sinais acústicos coletados de 10 atores (5 homens e 5 mulheres), em interações dois a dois. 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 foram: Raiva, Felicidade, Neutro e Tristeza. Para cada um desses 4 estilos, foi utilizado 10 minutos de trechos sonoros dos sinais de voz.

2) SEMAINE: Esta foi coletada a partir de dados audiovisuais de estudantes de graduação e pós gradução de oito diferentes 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 diferentes estilos, os quais consistiam de dimensões (ativação e valência, por exemplo) e emoções individuais, como Raiva e Tristeza. Neste trabalho, foram consideradas 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 classificação multiestilo. Para cada estado emocional, foi utilizado 90 segundos de trechos sonoros dos sinais de voz.

B. Extração dos Atributos Acústicos

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). Para a obtenção dos atributos MFCC e GFCC, os sinais de voz são segmentados em trechos de 25 ms, com uma taxa de quadro de 10 ms. Em relação ao MFCC, são obtidos 12 coeficientes, utilizando um banco com 20 filtros triangulares. Para os GFCCs, foram extraídos 22 coeficientes, utilizando um banco com 64 filtros Gammatone.

C. Classificadores

Para a tarefa de classificação dos múltiplos estados emocionais, foram empregados dois classificadores: GMM (*Gaussian Mixture Model*s) [30] e HMM (*Hidden Markov Models*) [31]. Os modelos emocionais GMM 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.

IV. Resultados

Os resultados dos experimentos realizados são apresentados nesta Seção, considerando as taxas de classificação obtidas para as bases IEMOCAP e SEMAINE, respectivamente.

A. Classificação com a IEMOCAP

Na Tabela I estão os valores percentuais de acurácia obtidos com os três atributos acústicos empregados neste trabalho, utilizando a base IEMOCAP. Note que para ambos os classificadores empregados nos experimentos, o vetor pH é o atributo que obtém os maiores valores de acurácia média. Como destaque, a abordagem GMM obtém, para o vetor pH, uma taxa de acerto de 51,0%, contra 49,3% obtido com HMM. Em relação aos atributos do trato vocal, o GFCC supera o MFCC com os dois classificadores. No contexto do GMM, o GFCC atinge uma acurácia média de 48,7% enquanto o MFCC chega a 47,5%. Todavia, vale ressaltar o potencial do vetor pH em identificar pelo menos três estados emocionais com mais de 50% de precisão. Na análise de cada emoção individualmente, a maior diferença de acurácia entre o vetor pH e os atributos do trato vocal, considerando o GMM, é para o estado Neutro, 5 pontos percentuais (p.p.). No contexto do HMM, a classificação da emoção Raiva foi o maior destaque do vetor pH em relação aos demais atributos, em que atinge, por exemplo, 5 p.p. a mais que o GFCC.

Na Figura 2 são apresentados os valores de classificação de cada estado emocional considerado da IEMOCAP, por meio da fusão dos três atributos acústicos utilizados nos experimentos.

TABELA I

TAXAS DE ACURÁCIA (%) UTILIZANDO ATRIBUTOS ACÚSTICOS PARA 4 ESTADOS EMOCIONAIS DA BASE IEMOCAP.

	Emoção	Emoção classificada com GMM				Emoção classificada com HMM			
	Real	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.
Vetor pH	Raiva	**57**	26	13	4	**57**	26	13	4
	Felicidade	30	**45**	17	8	33	**42**	17	8
	Neutro	12	15	**50**	23	12	15	**49**	24
	Tristeza	9	13	26	**52**	10	14	27	**49**
		Taxa de acurácia média: **51,0**				Taxa de acurácia média: **49,3**			
MFCC	Emoção	Emoção classificada com GMM				Emoção classificada com HMM			
	Real	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.
	Raiva	**54**	20	16	10	**50**	19	18	13
	Felicidade	31	**40**	20	9	30	**37**	22	11
	Neutro	16	11	**45**	28	16	12	**44**	28
	Tristeza	9	11	29	**51**	10	12	28	**50**
		Taxa de acurácia média: **47,5**				Taxa de acurácia média: **45,3**			
GFCC	Emoção	Emoção classificada com GMM				Emoção classificada com HMM			
	Real	Rai.	Fel.	Neu.	Tri.	Rai.	Fel.	Neu.	Tri.
	Raiva	**55**	19	16	10	**52**	18	17	13
	Felicidade	29	**43**	19	9	28	**39**	22	11
	Neutro	16	11	**45**	28	16	12	**45**	27
	Tristeza	9	11	28	**52**	10	12	27	**51**
		Taxa de acurácia média: **48,7**				Taxa de acurácia média: **46,7**			

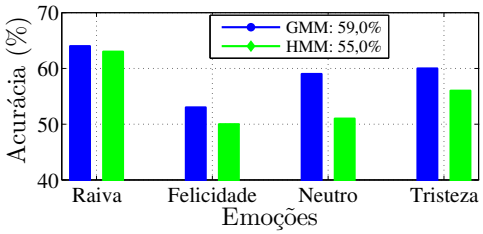

Fig. 2. Fusão dos atributos pH, MFCC e GFCC utilizando GMM e HMM para a base IEMOCAP.

TABELA II

TAXAS DE ACURÁCIA (%) UTILIZANDO ATRIBUTOS ACÚSTICOS PARA 4 ESTADOS EMOCIONAIS DA BASE SEMAINE.

	Emoção	Emoção classificada com GMM				Emoção classificada com HMM			
	Real	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.
Vetor pH	Raiva	**50**	22	20	8	**45**	25	22	8
	Felicidade	17	**51**	27	5	19	**47**	29	5
	Diversão	16	25	**49**	10	16	28	**45**	11
	Tristeza	9	16	24	**51**	8	18	27	**47**
		Taxa de acurácia média: **50,2**				Taxa de acurácia média: **46,0**			
MFCC	Emoção	Emoção classificada com GMM				Emoção classificada com HMM			
	Real	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.
	Raiva	**40**	30	17	13	**38**	31	17	14
	Felicidade	19	**49**	28	4	19	**49**	28	4
	Diversão	15	31	**46**	8	16	31	**42**	11
	Tristeza	11	12	26	**51**	10	13	30	**47**
		Taxa de acurácia média: **46,5**				Taxa de acurácia média: **44,0**			
GFCC	Emoção	Emoção classificada com GMM				Emoção classificada com HMM			
	Real	Rai.	Fel.	Div.	Tri.	Rai.	Fel.	Div.	Tri.
	Raiva	**45**	27	15	13	**41**	30	16	13
	Felicidade	19	**50**	27	4	20	**49**	27	4
	Diversão	14	31	**47**	8	15	33	**44**	8
	Tristeza	10	13	27	**50**	10	14	28	**48**
		Taxa de acurácia média: **48,0**				Taxa de acurácia média: **45,5**			

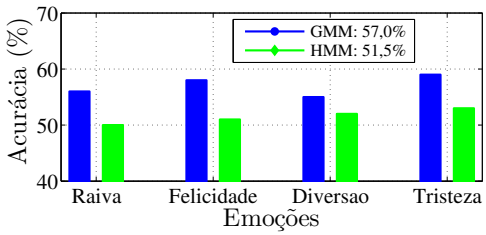

Fig. 3. Fusão dos atributos pH, MFCC e GFCC utilizando GMM e HMM para a base SEMAINE.

A maior acurácia é obtida utilizando o classificador GMM: 59,0%. Isto significa um aumento nas taxas obtidas por cada atributo individualmente. Com ambos os classificadores, a emoção Raiva é identificada com mais de 60% de precisão. Note que por meio da fusão de atributos, a classificação de cada estado emocional individualmente atinge no mínimo 50% de acurácia, tanto para GMM quanto para HMM. Os atributos do trato vocal são os que tem um maior aumento nos valores de acurácia, obtidos por meio da fusão com o vetor pH. A taxa média de acerto de MFCC e GFCC sobe 11,5 p.p. e 10,3 p.p., respectivamente, empregando GMM.

B. Classificação com a SEMAINE

Os resultados de classificação de múltiplos estados emocionais da base SEMAINE são apresentados na Tabela II. Para todos os atributos empregados, o classificador GMM superou HMM. O vetor pH foi o único atributo com o qual foi atingida uma acurácia média acima de 50,0%, o que é resultado de uma identificação de pelo menos 50,0% para três estados emocionais: Raiva, Felicidade e Tristeza. Em termos da taxa média de acerto, o vetor pH supera o MFCC em 3,7 p.p. e 2 p.p. com GMM e HMM, respectivamente. Em relação ao GFCC, o vetor pH atinge 2,2 p.p. a mais com GMM e 0,5 p.p. a mais com HMM. Note que, embora o atributo da fonte de excitação (vetor pH) supere aqueles do trato vocal, assim como ocorre para a base IEMOCAP, o GFCC supera o MFCC no contexto da SEMAINE. Individualmente, a classificação da emoção Raiva foi o maior destaque do vetor pH em relação a GFCC e MFCC, considerando ambos os classificadores. Por exemplo, a diferença entre vetor pH e GFCC, para este estado emocional, é de 5 p.p. e 4 p.p. utilizando GMM e HMM, respectivamente.

Os valores de acurácia obtidos por meio da fusão de atributos na base SEMAINE estão apresentados na Figura 3. Com ambos os classificadores empregados, a fusão do pH com MFCC e GFCC atinge mais de 50% de precisão média na identificação das variações acústicas emocionais. Como destaque, o GMM atingiu uma taxa média de acerto de 57%. Em relação à classificação de cada estado emocional com GMM, o maior ganho ocorre para o MFCC, que com a emoção Raiva vai de 40% a 56% quando agregado aos demais atributos. Para os outros estados emocionais, a melhora proporcionada por pH e GFCC ao MFCC é de 9 p.p., 9 p.p. e 8 p.p. com GMM para Felicidade, Diversão e Tristeza, respectivamente.

C. Discussão

A classificação de múltiplas emoções por meio de um atributo acústico que efetivamente caracterize cada estado emocional ainda é um desafio. Como foi observado nos resultados, apenas o vetor pH consegue uma acurácia média de no mínimo 50,0%, por meio da modelagem de cada um dos estados emocionais com misturas Gaussianas (GMM). Embora o GFCC tenha obtido um melhor desempenho em relação ao MFCC nesse contexto de variações acústicas emocionais, ambos foram superados pelo vetor pH. Isto é um indício de que atributos da fonte de excitação são mais propícios a este tipo de tarefa, uma vez que os atributos do trato vocal são mais

sensíveis ao contexto de fala e ao conteúdo linguístico [21]. No caso de uma busca por aumentar as taxas de acerto, pode ser considerada a fusão de atributos, o que pode levar à montagem de conjuntos com vários atributos [32]. Porém, mesclar uma grande quantidade de atributos acústicos seria mais uma tentativa de melhora de acurácia do que uma busca por uma impressão acústica emocional. Nos experimentos de fusão empregados neste trabalho, foi observado que há um aumento nas taxas de acerto, principalmente se forem observados os atributos do trato vocal. Isto é um indício de que o atributo da fonte de excitação (vetor pH) pode agregar informação consideravelmente útil a conjuntos de características que tenham como objeto a classificação de emoções.

V. Conclusão

Este trabalho apresentou a análise da classificação de múltiplas emoções a partir de duas bases acústicas, gravadas em diferentes contextos de comunicação entre os interlocutores. Para tanto, foram empregados um atributo da fonte de excitação glotal (vetor pH) e dois atributos do trato vocal (MFCC e GFCC). A classificação realizada com duas abordagens abordagens (GMM e HMM) mostrou que o vetor pH supera os demais atributos acústicos. Com o GMM, o vetor pH atinge mais de 50,0% de acurácia média na classificação de quatro estados emocionais nas duas bases de dados consideradas nos experimentos. Na fusão de atributos, o vetor pH proporcionou um considerável aumento nas taxas de acerto dos demais atributos, chegando a 59,0% de acurácia com a base IEMOCAP e a 57,0% com a base SEMAINE. Estes resultados indicam que atributos da fonte de excitação podem ser mais efetivos na caracterização e classificação de múltiplas variações acústicas emocionais.

Referências

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

[9] V. Vieira, R. Coelho, and F. M. de Assis, "Decomposição tempo-frequência adaptativa de sinais acústicos com variações emocionais," in *Anais do XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, pp. 65–69, 2016.

[10] V. Vieira, R. Coelho, and F. M. de Assis, "Análise de variações acústicas não estacionárias em sinais de voz gerados em condições de estresse," in *Anais do XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, pp. 737–741, 2017.

[11] 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.

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

[13] X. Zhao and D. Wang, "Analyzing noise robustness of mfcc and gfcc features in speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, pp. 7204–7208, IEEE, 2013.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.