Atributos Acústicos Baseados na Simetria Glotal e no Classificador α-GMM para Identificação de Emoções e Locutor

D. Cavalcante e R. Coelho

***Resumo*— Este trabalho apresenta um estudo sobre o reconhecimento acústico de emoções e seu efeito no desempenho de sistemas de reconhecimento de locutor. Experimentos de identificação de múltiplas emoções foram realizados utilizando um atributo baseado na simetria do pulso glotal (SG), estimada a partir da fonte de excitação do sinal de voz, e comparados com os resultados obtidos para o atributo CB-TEO-Auto-Env. Também foram efetuados testes de identificação conjunta de emoção e locutor através de fusão com os coeficientes Mel-cepestrais (MFCC). O efeito de degradação na acurácia permitiu a discriminação das emoções de acordo com o parâmetro α do classificador α-GMM.**

***Palavras-Chave*— reconhecimento automático de locutor, identificação, emoções primárias, TEO, simetria glotal, fonte de excitação, MFCC, α-GMM.**

***Abstract*— This paper presents a study of acoustic emotion recognition and its effects on speaker recognition systems. Multistyle emotion identification experiments were performed using a feature based on the glottal pulse symmtery, which is estimated from the source excitation signal, and compared with the results achieved by the CB-TEO-Auto-Env feature. Speaker and emotion dual identification experiments were also performed through feature fusion with Mel-frequency cepstral coefficients (MFCC). The degration on the accuracy allowed the discrimination of emotions along with the α value of the α-GMM classifier.**

***Keywords*— automatic speaker recognition, identification, primary emotions, TEO, glottal symmetry, source excitation, MFCC, α-GMM.**

I. Introdução

O sinal de voz transmite informações que podem ser agrupadas em dois canais de comunicação: um explícito, que carrega a mensagem a ser transmitida, e um implícito, que contém características do locutor. Através do canal implícito podem ser extraídas a identidade bem como o estado emocional ou nível de estresse a que o locutor está sujeito. Os estados emocionais podem ser identificados porque afetam os mecanismos de produção da fala, alterando consequentemente o sinal de voz [1], [2]. Os estados emocionais provocam alterações no sinal de voz que podem ser consideradas como variações ou oscilações acústicas. O reconhecimento acústico de emoções (RAE) visa classificar as emoções pelo sinal de voz. Sistemas de RAE têm recebido crescente interesse por ser considerado um método de classificação mais natural e menos intrusivo [3], [4].

Dirceu Cavalcante, Programa de Pós-graduação em Engenharia de Defesa , e Rosângela Coelho, Departamento de Engenharia Elétrica, Instituto Militar de Engenharia, Rio de Janeiro, Brasil, E-mails: dirceu_cavalcante@ime.eb.br, coelho@ime.eb.br.

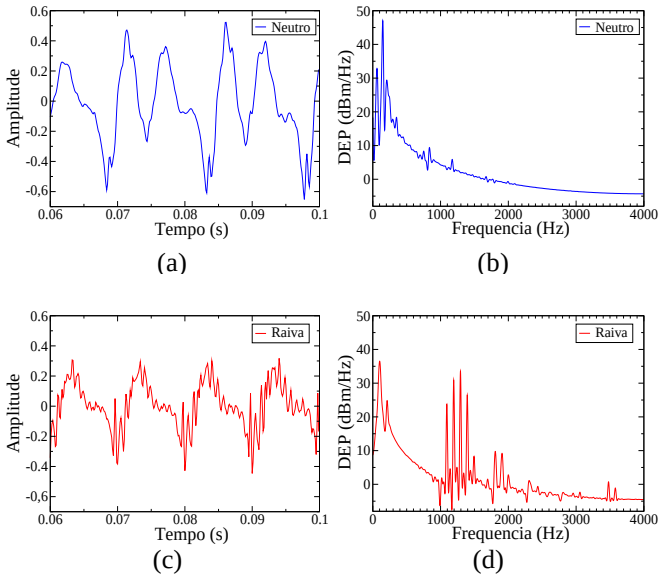

Fig. 1. Formas de onda do sinal de voz e suas correspondentes DEP de um locutor sob as emoções neutro (a,b) e raiva (c,d).

Dentre os diversos mapeamentos de grupos de emoções, destaca-se a separação em eixos, especialmente os eixos de ativação e valência [1], [3], [5]. O eixo de ativação está associado à mudanças fisiológicas, enquanto o eixo de valência está associado a experiência psicológica, sendo categorizada como positiva ou negativa [1]. No trato vocal, emoções induzem alterações na tensão muscular, na rigidez das cordas vocais e no fluxo de ar que sai dos pulmões. Portanto, a frequência fundamental e seus harmônicos carregam informações sobre as emoções. Para ilustrar tal efeito, a Fig. 1 exibe dois sinais de voz produzidos pelo mesmo locutor sob o estado neutro (a) e raiva (c). Perceba que na forma de onda produzida sob raiva há a presença de componentes de alta frequência, quando comparada a contraparte neutra. Pela densidade espectral de potência (DEP) destes segmentos, também é possível notar um deslocamento na energia dos harmônicos da pitch. Desta forma, informações extraídas a partir da fonte de excitação são capazes de discriminar emoções.

A extração de atributos capazes de identificar múltiplas emoções ainda é um problema em aberto [5]. Atributos espectrais, como o *Mel-frequency cepstral coefficients* (MFCC), prosódicos (envelope da pitch, por exemplo), de qualidade de voz (como a estrutura temporal de fonemas) ou mesmo baseados no operador de energia Teager (TEO) já foram analisados

para reconhecimento de emoções [5]. Há poucos estudos sobre atributos específicos para a classificação de múltiplas emoções, ou seja, atributos acústicos que permitam uma discriminação natural dos diferentes estados emocionais. O atributo *Critical-Band TEO Autocorrelation Envelope* (CB-TEO-Auto-Env) [2] foi utilizado como atributo de identificação de emoções em [6]. Neste trabalho, a simetria do pulso glotal [4] é analisada como atributo para representação e classificação de emoções através do sinal de voz. A simetria do pulso glotal (SG) é definida como a razão entre as fases de fechamento e abertura da glote. O sinal glotal é estimado a partir da fonte de excitação do sinal de voz, afetado fortemente pelo efeito de distorção emocional.

As alterações no aparelho fonador induzidas por estados emocionais afetam o desempenho de sistemas de reconhecimento automático de locutor (RAL) [6]. Aplicações forenses, por exemplo, são fortemente influenciadas por fatores temporais e emocionais [7], [8]. Portanto, o RAE visando o aumento de robustez de sistemas de RAL a estados emocionais é de grande relevância. Baseado neste cenário, o objetivo deste artigo é também de analisar a tarefa conjunta de identificação de emoções e de locutor. Um conjunto de 10 locutores foi utilizado sob 6 estados emocionais, extraídos da *Berlin Emotional Database* (EMO-DB) [9]. Para a identificação de múltiplas emoções, foram analisados o desempenho dos atributos SG e CB-TEO-Auto-Env em conjunto com o classificador de Misturas Gaussianas α-integráveis (α-Gaussian Mixture Models) [10]. O α-GMM foi proposto para minimizar efeitos de descasamento entre amostras de teste e treinamento através da escolha do parâmetro α, oriundo do conceito de α-integração. Cada valor de α representa uma família de distintos classificadores GMM. Quando $\alpha = -1$, o α-GMM é o classificador GMM convencional. Experimentos de identificação de locutor foram realizados utilizando o atributo MFCC para um sistema α-GMM treinado apenas com locuções no estado neutro. A dupla tarefa de identificação de locutor e emoção foi realizada através da fusão dos atributos MFCC e CB-TEO-Auto-Env para diversas configurações do classificador α-GMM.

Este artigo é dividido em seis seções. Na Seção II são descritos os atributos acústicos SG e CB-TEO-Auto-Env. O classificador α-GMM é apresentado na Seção III. Os resultados dos experimentos de identificação de locutor e de emoções são mostrados na Seção IV, juntamente com uma breve descrição da base utilizada. Por fim, as conclusões são apresentadas na Seção V.

II. Atributos para a identificação de emoções

Nesta Seção são apresentados os atributos acústicos utilizados para o RAE. O atributo SG discrimina as emoções pela variação do formato do pulso glotal, através da razão entre os instantes de fechamento e abertura da glote. O atributo CB-TEO-Auto-Env, por sua vez, mapeia as variações na distribuição dos harmônicos da pitch em faixas de frequência da audição humana.

A. Atributo Simetria Glotal (SG)

Em sinais de voz sonoros, o ar passa através da glote gerando um fluxo pulsado que se propaga no trato vocal como uma onda acústica plana. Este fluxo pulsado é resultante da vibração das cordas vocais, no período fundamental T. Do modelo de produção linear da fala [11], o sinal de voz resultante, $S(z)$, medido nos lábios, pode ser expresso como

$$S(z) = G(z)V(z)R(z), \tag{1}$$

onde $G(z)$ e $V(z)$ são o sinal de excitação glotal e o modelo do trato vocal, respectivamente, e $R(z)$ é o efeito de radiação nos lábios. O sinal glotal pode ser estimado por filtragem inversa de (1)

$$G(Z)R(z) = \frac{S(z)}{V(z)}, \tag{2}$$

onde $V(z)$ é um filtro apenas com pólos, cujos coeficientes são obtidos a partir da análise de predição linear (LP) durante os instantes de tempo em que a glote permanece fechada [12]. Nesta aproximação, o sinal de voz pode ser considerado uma oscilação desvanecente devido a ressonância nas paredes do trato vocal. O efeito de radiação pode ser removido através da integração do sinal glotal.

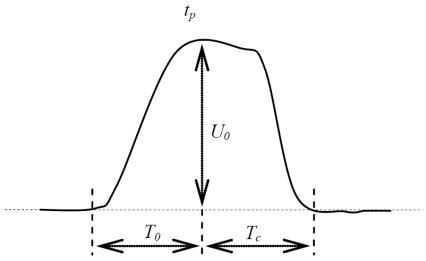

Fig. 2. Modelo de pulso glotal proposto por Fant [13].

Os instantes de fechamento da glote (*Glotal Closure Instante* - CGI) são identificados aplicando-se a função de atraso de grupo ponderada no espectro (*energy-weighted group delay function* - d_{EW}),

$$d_{EW} = \frac{\sum_{n=0}^{N-1} n x_r^2(n)}{\sum_{n=0}^{N-1} x_r^2(n)}, \tag{3}$$

onde $x_r(n)$ é um segmento de voz sonoro janelado de comprimento N começando na amostra r. Os CGIs correspondem então aos cruzamentos negativos por zero. A janela desta análise foi igual a 20 ms.

Um novo sinal, composto das amostras correspondentes à fase de fechamento da glote, é utilizado em uma nova análise LP do trato vocal. O sinal residual resultante é então integrado, obtendo-se o sinal glotal. O modelo de Fant [13] para o pulso glotal foi adotado e é ilustrado na Fig. 2. Os pontos de CGI no sinal glotal correspondem ao pico do pulso glotal (t_p). A simetria glotal é então dada por $SG = T_c/T_o$, onde T_c e T_o correspondem aos períodos de fechamento e abertura da glote.

B. Atributo CB-TEO-Auto-Env

A motivação deste atributo, proposto em [2], é capturar informação dependente de emoção que podem estar presentes em variações na componente FM do sinal de voz. Baseado no fato que o sistema auditório humano é capaz de realizar

operações de filtragem em partições da faixa de frequência audível, variações no padrão de modulação podem ser obtidas através da aplicação do TEO em sinais filtrados nessas bandas críticas. Nestas condições, o sinal de voz pode ser considerado um sinal AM-FM:

$$r(n) = a(n)cos\left[nw(n)\right]. \quad (4)$$

Assumindo-se que existam apenas duas harmônicas ω_{η_1} e ω_{η_2} na saída $\eta^i(n)$ de uma particular banda crítica i sobre condições neutras, enquanto exista apenas uma harmônica ω_{ν_1} na saída $\nu^i(n)$ da mesma banda crítica. As representações AM-FM pode ser explicitadas como:

$$\eta^i(n) = A_{\eta_1} cos\left[\omega_{\eta_1} n\right] + A_{\eta_2} cos\left[\omega_{\eta_2} n\right], \quad (5)$$

$$\nu^i(n) = A_{\nu_1} cos\left[\omega_{\nu_1} n\right], \quad (6)$$

onde as componentes AM e FM são consideradas constantes. Aplicando-se o operador TEO nos sinais $\omega^i(n)$ e $\nu^i(n)$, temos como resultado:

$$\Psi\left[\eta^i(n)\right] = A^2{}_{\eta_1} sen^2\left[\omega_{\eta_1}\right] + A^2{}_{\eta_2} sen^2\left[\omega_{\eta_2}\right] + 2A_{\eta_1}A_{\eta_2}\Big(sen^2\left[\frac{\omega_{\eta_1}+\omega_{\eta_2}}{2}\right] cos\left[(\omega_{\eta_1}-\omega_{\eta_2})n\right] + sen^2\left[\frac{\omega_{\eta_1}-\omega_{\eta_2}}{2}\right] cos\left[(\omega_{\eta_1}+\omega_{\eta_2})\,n\right]\Big), \quad (7)$$

$$\Psi\left[\nu^i(n)\right] = A^2{}_{\nu_1} sen^2\left[\omega_{\nu_1}\right]. \quad (8)$$

Percebe-se que o resultado do operador TEO no sinal sob emoção é uma constante, enquanto no sinal neutro é uma função do tempo consistindo de duas frequências $\omega_{\eta_1}+\omega_{\eta_2}$ e $|\omega_{\eta_1}-\omega_{\eta_2}|$.

Como a saída de uma banda crítica pode possuir harmônicos cruzados além de múltiplos da frequência fundamental e os termos AM e FM podem variar com o tempo, o cálculo da área sob a função de autocorrelação normalizada pode suprimir variações rápidas, mantendo as informações sobre as variações na componente FM intactas.

III. Classificador α-GMM

O GMM conventional ($\lambda_{\mathcal{E}}$) de um locutor sob um estado emocional $\mathcal{E}$ é definido como uma combinação linear de M componentes Gaussianas,

$$p(\vec{x}|\lambda_{\mathcal{E}}) = \sum_{j=1}^{M} p_j b_j(\vec{x}), \quad (9)$$

onde $\vec{x}$ é um vetor de atributos de voz de dimensão D, p_j são os pesos da mistura, onde $\sum_{j=1}^{M} p_j = 1$, e $b_j(\vec{x})$ são densidades Gaussianas com vetores média $\vec{\mu_j}$ e matrizes covariância K_j, ou seja,

$$b_j(\vec{x}) = \frac{\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu_j})K_j^{-1}(\vec{x}-\vec{\mu_j})\right)}{(2\pi)^{\frac{D}{2}}\sqrt{det(K_j)}}. \quad (10)$$

Então, o GMM de um locutor sob um estado emocional $\mathcal{E}$ pode ser completamente parametrizado por

$$\lambda_{\mathcal{E}} = \{p_j, \vec{\mu_j}, K_j | j = 1, \ldots, M\}. \quad (11)$$

Usando o conceito de α-integração [10], o α-GMM pode ser definido como

$$p_\alpha(\vec{x}|\lambda_{\mathcal{E}}) = \begin{cases} c \cdot \left(\sum_{j=1}^{M} p_j \left(b_j(\vec{x})\right)^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} & , \alpha \neq 1 \\ c \cdot exp\left(\sum_{j=1}^{M} p_j \log\left(b_j(\vec{x})\right)\right) & , \alpha = 1 \end{cases}, \quad (12)$$

onde c é uma constante tornar $p_\alpha(\vec{x}|\lambda_{\mathcal{E}})$ uma função densidade de probabilidade.

Note que quando $\alpha = -1$, (12) é o GMM convencional como definido em (9). Uma característica interessante do α-GMM é que usando valores negativos de α, o α-GMM diminui a ênfase nas baixas probabilidades enquanto enfatiza as maiores probabilidades, minimizando assim o descasamento entre as amostras de teste e treinamento. O α-GMM de um locutor sob um estado emocional pode também ser parametrizado por (11).

IV. Resultados

Nesta Seção são apresentados o conjunto experimental utilizado e a análise dos resultados. A tarefa de identificação de emoções visa estudar a capacidade dos atributos acústicos em discriminar múltiplos estados emocionais. A capacidade do classificador α-GMM em diminuir o descasamento entre as amostras de teste e treino é analisada nos experimentos de identificação de locutor sob variação acústica emocional. Dado que a informação sobre as emoções é dependente do locutor, testes de identificação conjunta de locutor e emoções são realizados para diferentes valores do parâmetro α.

A. Base de Dados

Para os testes de identificação de emoções e locutor, utilizou-se a base de dados pública *Berlin Emotional Database* (EMO-DB) [9]. Dez atores (5 homens e 5 mulheres) simularam seis emoções (raiva, felicidade, tédio, medo, neutro, tristeza e repugnância), produzindo sentenças curtas e longas, num total de 800 locuções. As gravações foram realizadas na câmara anecóica da *Technical University Berlin* a uma frequência de amostragem de 48 kHz e posteriormente subamostradas a 16 kHz. Para garantir a naturalidade e qualidade emocional, realizou-se um teste subjetivo, reduzindo-se o conjunto para 535 locuções.

Nos experimentos, as locuções foram subamostradas a 8 kHz e o silêncio foi removido. Um sistema de *round-robin* foi montado com 2044 segmentos de 1 s divididos aleatoriamente em 4 grupos. A cada bateria de testes, 3 grupos foram utilizados para treinamento. Desta forma, toda a base foi utilizada para treinamento e testes.

O classificador α-GMM foi configurado para valores do parâmetro α na faixa de -1 a -8. Os modelos foram gerados com ordem 16, onde cada locutor possui 6 modelos referentes aos estados emocionais.

Para aplicação em sistemas de RAL, uma adaptação na extração do CB-TEO-Auto-Env foi efetuada para desconsiderar a dependência da informação fonética (apenas sinais de

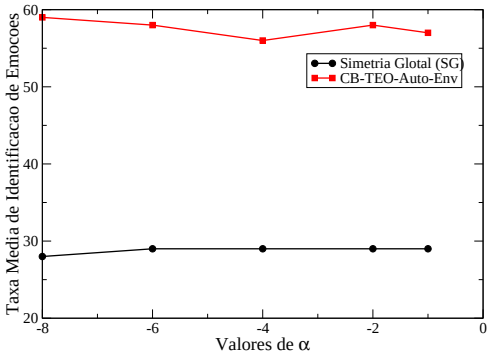

Fig. 3. Taxa média de identificação de múltiplas emoções dos atributos Simetria Glotal (SG) e CB-TEO-Auto-Env utilizando várias configurações do classificador α-GMM.

TABELA I

TAXA DE ACERTO DE IDENTIFICAÇÃO DE MÚLTIPLAS EMOÇÕES (%) UTILIZANDO SG PARA TESTES DE 1 S E $\alpha = -2$

Emoção Real	Emoção Reconhecida					
	Raiva	Tédio	Medo	Felicidade	Neutro	Tristeza
Raiva	**36,62**	14,04	12,71	16,51	12,90	7,21
Tédio	18,97	**35,63**	11,49	8,05	16,95	8,91
Medo	26,58	22,36	**21,52**	8,86	14,35	6,33
Felicidade	31,44	16,29	11,74	**24,62**	11,36	4,55
Neutro	19,32	18,64	8,14	8,14	**36,27**	9,49
Tristeza	14,04	15,20	13,16	13,45	20,47	**23,68**

TABELA II

TAXA DE ACERTO DE IDENTIFICAÇÃO DE MÚLTIPLAS EMOÇÕES (%) UTILIZANDO TEO-CB-AUTO-ENV PARA TESTES DE 1 S E $\alpha = -8$

Emoção Real	Emoção Reconhecida					
	Raiva	Tédio	Medo	Felicidade	Neutro	Tristeza
Raiva	**76,36**	1,89	0,95	17,05	0,57	0,19
Tédio	2,81	**58,99**	3,93	3,93	18,82	11,52
Medo	10,97	15,19	**38,40**	14,35	7,17	13,92
Felicidade	32,95	8,71	2,65	**53,79**	0,76	1,14
Neutro	3,72	28,72	2,36	3,04	**55,74**	6,42
Tristeza	1,39	14,72	3,89	1,11	7,22	**71,67**

voz sonoros de alta energia) e visando o acoplamento com o MFCC [6]. O sinal de voz é dividido em quadros utilizando-se a janela de Hamming, com duração de 20 ms e sobreposição de 50%, seguido da aplicação em um banco de filtros de Gabor contendo 16 filtros centrados em bandas críticas. O TEO é aplicado em cada um desses sinais, e a área normalizada da autocorrelação é então calculada. A dimensão do vetor de atributos é então igual ao número de filtros de Gabor.

B. Identificação de Emoções

Como a informação sobre a distorção acústica emocional é encontrada nos sons sonoros, os sons surdos foram removidos de cada segmento nos experimentos.

O vetor de atributos SG foi extraído a partir de janelas de 50 ms de cada segmento. Cada vetor contém 5 valores de simetria glotal consecutivos. O CB-TEO-Auto-Env foi extraído em quadros de 20 ms com 50% de sobreposição.

A Fig. 3 ilustra a variação da taxa média de identificação de emoções para várias configurações do classificador α-GMM. Cada atributo apresentou um valor diferente do parâmetro α para a maior taxa média de identificação obtida, sendo o SG menos sensível a esta variação. O atributo CB-TEO-Auto-Env apresentou os melhores resultados. O atributo SG teve o desempenho afetado pelo tamanho da população e tipos de emoção e fonemas analisados.

TABELA III

TAXA DE ACERTO DE DISCRIMINAÇÃO DE EMOÇÕES (%) QUANTO AOS EIXOS DE ATIVAÇÃO E VALÊNCIA PARA O ATRIBUTO SG

Emoção	Ativação		Valência	
	Alta	Baixa	Positiva	Negativa
Raiva	**65,84**	34,16	29,42	**70,58**
Tédio	38,51	**61,49**	25,00	**75,00**
Medo	**56,96**	43,04	23,21	**76,79**
Felicidade	**67,80**	32,20	**52,58**	47,42
Neutro	35,60	**64,50**	**44,41**	55,59
Tristeza	40,65	**59,35**	33,92	**59,35**

TABELA IV

TAXA DE ACERTO DE DISCRIMINAÇÃO DE EMOÇÕES (%) QUANTO AOS EIXOS DE ATIVAÇÃO E VALÊNCIA PARA O ATRIBUTO CB-TEO-AUTO-ENV

Emoção	Ativação		Valência	
	Alta	Baixa	Positiva	Negativa
Raiva	**94,36**	5,64	20,61	**79,39**
Tédio	10,67	**89,33**	22,75	**77,25**
Medo	**63,72**	36,28	21,52	**78,48**
Felicidade	**89,39**	10,61	**54,55**	45,45
Neutro	9,12	**90,88**	**58,78**	41,22
Tristeza	6,39	**93,61**	8,33	**91,67**

TABELA V

TAXA DE ACERTO DA IDENTIFICAÇÃO DE LOCUTOR (%) UTILIZANDO MFCC, TREINADO COM LOCUÇÕES NO ESTADO NEUTRO, TESTES DE 1 S E $-1 \leq \alpha \leq -8$.

Emoção	$\alpha = -1$	$\alpha = -2$	$\alpha = -4$	$\alpha = -6$	$\alpha = -8$
Raiva	12,69	13,45	13,26	13,07	12,88
Tédio	69,38	69,38	69,94	69,66	69,38
Medo	27,85	29,11	28,29	29,11	27,85
Felicidade	18,56	21,59	19,70	18,56	20,08
Neutro	99,66	99,99	99,66	99,99	99,32
Tristeza	75,56	73,89	77,22	76,67	75,56
Média	**50,62**	**51,24**	**51,41**	**51,18**	**50,84**

Nas Tabelas I e II são apresentadas as matrizes de confusão correspondentes às maiores taxas médias obtidas para a melhor configuração do classificador α-GMM e os atributos SG e CB-TEO-Auto-Env, respectivamente. Pelos resultados obtidos, o atributo SG não consegue distinguir as emoções medo e felicidade, sendo estas confundidas com raiva. O atributo baseado no TEO conseguiu classificar corretamente todas as emoções.

A acurácia da discriminação de emoções quanto aos eixos de ativação e valência é mostrada nas Tabelas III e IV para os atributos SG e CB-TEO-Auto-Env. Quanto aos eixos de ativação e valência, apenas o atributo baseado no TEO classificou corretamente todas as emoções. O atributo SG não discriminou apenas o estado neutro quanto ao eixo de valência.

C. Identificação de Locutor sob Estados Emocionais

Para os testes de identificação de locutor, foram utilizados 16 coeficientes MFCC obtidos a partir de janelas de 20 ms com

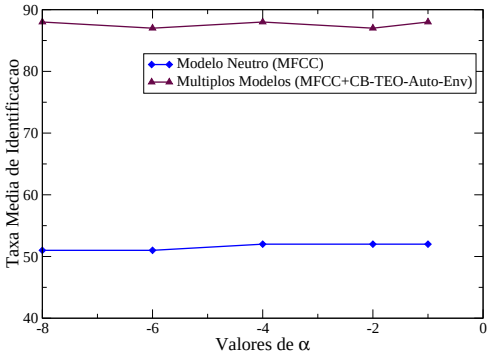

Fig. 4. Taxa média de identificação de Locutor sob múltiplas emoções com os atributos MFCC (Modelo Neutro) e MFCC+CB-TEO-Auto-Env (Múltiplos Modelos), para testes de 1 s, utilizando várias configurações do classificador α-GMM.

TABELA VI
TAXA DE ACERTO DA IDENTIFICAÇÃO DE LOCUTOR E EMOÇÃO (%) UTILIZANDO MFCC+TEO-CB-AUTO-ENV, TESTES DE 1 S E $\alpha = -4$

Emoção	Múltiplos Modelos
Raiva	92,23
Tédio	80,06
Medo	91,98
Felicidade	82,20
Neutro	93,24
Tristeza	94,17
Média	**88,98**

sobreposição de 50%. Os modelos α-GMM foram gerados com ordem 16. Apenas as locuções em estado neutro foram utilizadas para a geração de modelos na fase de treinamento.

A Tabela V exibe a acurácia do sistema de RAL submetido ao efeito de distorção emocional para diferentes valores do parâmetro α. Observa-se que o melhor desempenho médio é obtido para o valor $\alpha = -4$ (51,41%). De forma geral, a variação do valor de α ajuda a diminuir sutilmente o descasamento entre as condições de treinamento e teste.

A discriminação das emoções pode ser realizada quanto à degradação na acurácia obtida para o sistema de RAL. As emoções de alta ativação degradam fortemente a taxa de identificação. Isto se deve ao fato de que a energia é mais concentrada nas altas frequências, comparando-se com o estado neutro. O valor do parâmetro α também pode ser usado para classificar os estados emocionais. As emoções de alta ativação obtiveram as melhores taxas para $\alpha = -2$, enquanto as de baixa ativação, para $\alpha = -4$.

D. Identificação de Locutor e Emoção

Para a tarefa conjunta de identificação de locutor e emoção, utilizou-se vetores de atributos contendo 32 coeficientes (16 MFCC + 16 TEO-CB-AUTO-ENV) juntamente com modelos α-GMM de ordem 16. Cada locutor é associado então a 6 modelos. A taxa de identificação é definida como a razão entre os acertos do locutor e seu estado emocional pelo número total de testes.

A Fig. 4 ilustra uma comparação da variação da taxa média entre as tarefas de identificação de locutor e identificação conjunta com emoção para diversos valores do parâmetro α. Para ambos os casos, a maior acurácia é obtida para $\alpha = -4$. Observa-se que não há variação significativa na taxa média.

A Tabela VI exibe os resultados para a maior taxa média obtida para a identificação conjunta. As emoções felicidade e tédio obtiveram as menores acurácias: 82,20% e 80,06%, respectivamente. As emoções de alta ativação obtiveram melhor desempenho para $-1 \leq \alpha \leq -4$, enquanto as de baixa, para $-6 \leq \alpha \leq -8$.

V. CONCLUSÕES

Este artigo apresentou uma análise dos atributos CB-TEO-Auto-Env e SG para o reconhecimento acústico de emoções. O atributo baseado no TEO apresentou desempenho superior ao SG para as tarefas de identificação múltipla e classificação quanto aos eixos de ativação e valência. A tarefa conjunta de identificação de locutor e emoção foi analisada e obteve taxa média de 88,98%. A redução no descasamento para a tarefa de identificação de locutor, proposta pelo classificador α-GMM, treinado apenas com locuções neutras, foi sutil quando comparada ao GMM convencional ($\alpha = -1$). No entanto, a degradação na acurácia permitiu a separação das emoções quanto ao eixo de ativação. Também foi possível discriminar as emoções pelo parâmetro α. Os estados de alta ativação obtiveram melhor desempenho para $-1 \leq \alpha \leq -4$, enquanto os de baixa ativação para $-6 \leq \alpha \leq -8$.

REFERÊNCIAS

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, and W. Fellenz, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, n. 1, pp. 32–80, jan 2001.

[2] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, n. 3, pp. 201–216, mar 2001.

[3] 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 frist challenge," *Speech Communication*, pp. 1062–1087, 2011.

[4] A. Iliev and M. Scordilis, "Spoken emotion recognition using glottal symmetry," *EURASIP Journal on Advances in Signal Processing*, 2011.

[5] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech recognition: resources, features and methods," *Pattern Recognition*, vol. 44, n. 3, pp. 572–587, mar 2011.

[6] D. Cavalcante and R. Coelho, "Identificação de emoções aplicada ao reconhecimento automático de locutor," *Anais do XXIX Simpósio Brasileiro de Telecomunicações - SBrT'11*, Outubro 2011.

[7] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meigner, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delecrétraz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, pp. 431–451, Apr 2004.

[8] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, pp. 95–103, March 2009.

[9] F. Burkhardt, A. Paetchke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," *Proceedings of the Interspeech*, pp. 1517–1520, 2005.

[10] 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.

[11] L. Rabiner and R. Schafer, *Digital Procesing of Speech Signals*. Prentice Hall, 1978.

[12] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Transactions on Audio, Speech and Language*, vol. 14, n. 2, pp. 456–465, mar 2006.

[13] G. Fant, "Glottal source and excitation analysis," *Speech Transmission Laboratory, Quarterly Progress and Status Report*, no. 1, pp. 70–85, 1979.