Identificação de Emoções Aplicada ao Reconhecimento Automático de Locutor

D. Cavalcante e R. Coelho

***Resumo*— Este trabalho apresenta um estudo sobre o efeito das emoções no desempenho de sistemas de reconhecimento automático de locutor. Experimentos de identificação de emoções foram realizados utilizando um atributo baseado no operador de energia Teager (TEO) e o classificador de mistura Gaussiana (GMM). Os resultados mostram que a presença de emoções afeta fortemente a taxa de acertos na identificação de locutor. Uma solução baseada no operador TEO é também proposta para tornar a identificação de locutor robusta à presença de diferentes estados emocionais.**

***Palavras-Chave*— reconhecimento automático de locutor, identificação, emoções primárias, TEO, MFCC, GMM.**

***Abstract*— This paper presents an study of automatic speaker recognition systems performance in the presence of emotions. Experiments to identify the emotions were performed using an attribute based on Teager energy operator (TEO) and Gaussian mixture models (GMM) classifier. The results show that the presence of emotional states strongly affects the accuracy rate of speaker identification. A solution based on TEO is also proposed to make the speaker identification robust to the presence of different emotional states.**

***Keywords*— automatic speaker recognition, identification, primary emotions, TEO, MFCC, GMM.**

I. Introdução

A voz é um sinal acústico proveniente do sistema de produção de fala que contém características fisiológicas e comportamentais. As informações sobre a identidade do locutor, baseadas na estrutura fisiológica do trato vocal e seus anexos, são invariantes. Já as comportamentais variam com o tempo devido a fatores como idade, condições de saúde, estado emocional, nível de estresse, entre outros [1]. Associado ao fato que a voz é um sinal biométrico de fácil aquisição, sistemas de reconhecimento automático de locutor (RAL) [2], [3] têm ampla aceitação em diversas aplicações, tais como: segurança de informações, controle de acesso e defesa (investigações forenses e aplicações militares).

O desempenho de sistemas de RAL é fortemente degradado na presença de condições adversas de captação. Condições como a presença de ruídos acústicos ou o canal de transmissão impõem um alto grau de variabilidade no sinal de voz. Em [4], propôs-se uma classificação para a variabilidade da fala: intralocutor peculiar (gênero, separação temporal, estresse induzido por estado emocional), intralocutor forçada (estresse induzido por carga de trabalho [5] e efeito *Lombard* [6], [7]) e variação induzida por fonte externa (tipo de microfone, canal de comunicação, ruído acústico ambiental).

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

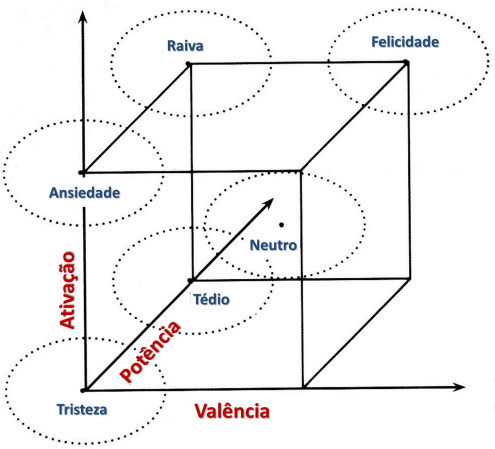

Fig. 1. Separação das emoções primárias (Raiva, Felicidade, Ansiedade, Neutro, Tédio e Tristeza) nos eixos valência, ativação e potência [8].

A robustez em sistemas de reconhecimento baseados no sinal de voz pode, portanto, abranger uma vasta gama de problemas. Tipicamente, a presença de ruídos acústicos, estresse ou distorção acarretam em forte queda no desempenho destes sistemas. Estresse foi definido em [7] como sendo qualquer condição que altere o processo de produção da fala. Portanto, a ausência de estresse caracterizaria um estado Neutro ou sem emoção (sinal de voz limpo). O estresse pode ser classificado como: mecânico (vibração ou aceleração), acústico (eco, ruído acústico ambiental) ou emocional. O efeito *Lombard*, por exemplo, é uma reação natural do locutor a ambiente ruidoso na tentativa de tornar melhorar a qualidade da fala, sendo então classificado como estresse acústico. A análise de robustez de sistemas de RAL na presença de ruidos acústicos frequentemente desconsidera tal efeito, pois o ruído é adicionado eletronicamente nas locuções.

O reconhecimento de emoções pela voz (REV) é uma tarefa de reconhecimento de padrões desafiadora por diversos motivos [9]. Por exemplo, não há clareza quanto ao tipo de atributo mais adequado à discriminação de emoções. Além disso, um estado emocional pode se prolongar por dias, semanas, anos ou mesmo períodos curtos [10]. Portanto, não existe uma definição se o sistema irá detectar um estado emocional longo ou emoção transiente. Outro problema é que não existe um consenso sobre a modelagem de emoções. Segundo a teoria da palheta [9], emoções podem ser decompostas em combinações de emoções primárias (Raiva, Felicidade, Repugnância, Medo, Tristeza, Neutro e Ansiedade). Outra abordagem difundida é a separação das emoções em em duas ou mais dimensões. A Fig. 1 ilustra uma representação da separação das emoções primarias em três eixos. Pode-se perceber que emoções como Felicidade e Raiva, por exemplo, são distintas apenas no eixo

de valência [9]. Estes eixos (valência, ativação e potência) [10], [8] são oriundos de uma abordagem psicológica e refletem alterações no sistema parassimpático durante estados emocionais. O eixo de ativação está associado à disposição de ação do indivíduo e se reflete no sinal de voz como variações na *pitch*. O eixo de valência está relacionado à noção de valor da experiência (positiva ou negativa), enquanto o eixo de potência é associado a energia do sinal de voz.

A classificação entre estados Neutro e Estressado já obtém boas taxas de reconhecimento de emoções utilizando atributos baseados no operador TEO [11]. No entanto, a identificação das emoções primárias ainda representa um grande desafio [9]. O operador TEO é baseado em observações sobre o modelo de produção não-linear da fala [12] e representa variações na energia do sinal de voz devido a interações vortex-fluxo no trato vocal.

O objetivo deste trabalho é analisar o desempenho de um sistema de RAL com locuções na presença de emoções primárias. Para a tarefa de identificação de emoções, utilizou-se um atributo baseada no operador TEO e o classificador GMM. Tal atributo é recomendado para classificação de Estresse [10] em sistemas de REV. Experimentos de identificação de emoções primárias dependentes de locutor e independentes de texto foram realizados num contexto de independência fonética. Resultados mostraram que a presença de emoções afeta o desempenho de sistemas de RAL no estado da arte. Os coeficientes MFCC [13] e o classificador GMM [14] foram utilizados nos experimentos de identificação de locutor. As locuções em estado Neutro (locução limpa) foram utilizadas para o treinamento. Por fim, verificou-se a contribuição da adição do atributo baseado no operador TEO na matriz de coeficientes MFCC no desempenho de um sistema de RAL no estado da arte.

Este artigo é organizado da seguinte forma. Na Seção II, as aproximações para extração de atributos empregados no REV são descritas, seguido pela descrição do operador TEO e a metodologia de extração adotada. O sistema de RAL no estado da arte é apresentado na Seção III, utilizando os coeficientes MFCC e o classificador GMM. 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. Extração de Atributos para Classificação de Emoções

Embora diversos métodos de extração de atributos tenham sido testados em sistemas de REV, ainda não há uma melhor aproximação para o problema. Em [10], estes atributos foram agrupados em quatro tipos:

1) *Prosódicos contínuos*: Relacionadas a *pitch* e formantes, energia, taxa de articulação e cadência;
2) *Qualitativos*: Relacionadas a qualidade da voz, como, por exemplo, limites de fonemas ou palavras, estruturas temporais e amplitude;
3) *Espectrais*: Baseadas na representação espectral da representação de tempo-curto do sinal de fala, como, por exemplo, MFCC e os coeficientes de predição linear (LPC);
4) *Baseados no operador TEO*: Relacionadas com a evidência que a audição é um processo de detectar energia.

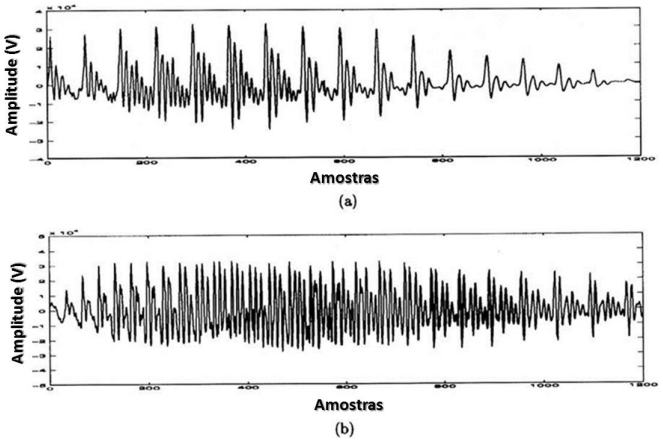

Fig. 2. Formas de onda do sinal de voz obtidas a partir da parte sonora da palavra inglesa "*Help*" faladas pelo mesmo locutor masculino sobre condições (a) Neutro e (b) estresse (Raiva simulada). Extraído de [11].

Para a tarefa de detecção de estresse na voz, atributos baseados no operador TEO apresentadas em [11] apresentaram resultados superiores ao MFCC e a *Pitch*. Atributos como MFCC e *Pitch* são baseados no modelo de produção linear da fala que assume que o fluxo de ar propagando-se no trato vocal como uma onda acústica plana é a fonte da produção de som. Estudos efetuados por Teager [12] mostraram que a verdadeira fonte da produção são as interações não-lineares do fluxo-vórtices dentro do trato vocal. Baseados nesta observação, acredita-se que as mudanças fisiológicas no trato vocal durante condições de estresse, como por exemplo tensão muscular, irão afetar os padrões de interação fluxo-vórtices [11]. Desta forma, atributos não-lineares devem ser capazes de melhor detectar a presença de estresse no sinal de voz. Na Fig. 2 pode-se notar as diferenças na *pitch* entre os sinais sonoros da palavra inglesa "*help*" do mesmo locutor sobre as condições Neutro e estresse por estado emocional Raiva.

A. Operador de Energia Teager

Na tentativa de refletir a energia das interações fluxo-vórtices, Teager desenvolveu um operador de energia baseado na observação que o processo de audição é um processo de detecção de energia. A expressão matemática do operador TEO proposto por Kaiser [11] é definido por:

$$\Psi\left[x(n)\right] = x^2(n) - x(n-1)x(n+1) \tag{1}$$

onde $\Psi\left[.\right]$ é o operador TEO e $x(n)$ é o sinal de voz amostrado.

O operador TEO é geralmente aplicado a um sinal de voz filtrado por um filtro passa-banda. A intenção do operador TEO é refletir a energia das interações não-lineares dentro do trato vocal para uma única frequência de ressonância. Embora no sinal filtrado ainda exista mais de uma frequência ressonante, o sinal de voz pode ser considerado como um sinal AM-FM,

$$r(n) = a(n)cos\left[nw(n)\right] \tag{2}$$

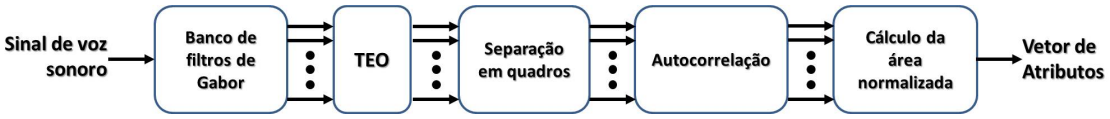

Fig. 3. Extração do atributo TEO-CB-AUTO-ENV como proposto em [11].

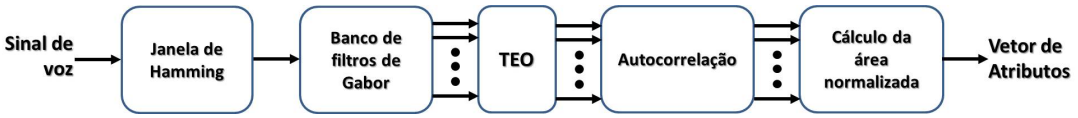

Fig. 4. Adaptação do atributo TEO-CB-AUTO-ENV para aplicação em reconhecimento de locutor.

onde $a(n)$ e $w(n)$ são as componentes AM e FM respectivamente. A aplicação do operador TEO no sinal de voz $r(n)$ pode ser aproximada por:

$$\Psi\left[r(n)\right] \approx \left[a(n)w(n)\right]^2. \tag{3}$$

O sinal resultante após aplicação do operador TEO, em um sinal que possui várias frequências ressonantes, reflete não apenas os componentes individuais como também as suas interações [11]. Como a presença de estresse na voz altera a frequência fundamental e os padrões de ressonância no trato vocal, atributos que representem finas variações nas frequências de excitação devem ser úteis para a classificação de estresse.

B. Extração do atributo TEO-CB-AUTO-ENV

A motivação para a extração do atributo TEO-CB-AUTO-ENV (*Critical Band based TEO Autocorrelation Envelope*), proposto em [11], é capturar a informação dependente de estresse que pode estar presente 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 operador TEO em sinais filtrados nessas bandas críticas.

Assumindo-se que existam apenas duas frequências harmônicas ω_{η_1} e ω_{η_2} na saída $\eta^i(n)$ de uma banda crítica i sobre condições neutras. Supondo-se a presença de apenas uma frequência harmônica ω_{ν_1} na saída $\nu^i(n)$ da mesma banda crítica para uma condição de estresse na voz. Desta forma, as representações AM-FM dos sinais $\eta^i(n)$ e $\nu^i(n)$ podem ser explicitadas como:

$$\eta^i(n) = A_{\eta_1} cos\left[\omega_{\eta_1} n\right] + A_{\eta_2} cos\left[\omega_{\eta_2} n\right] \tag{4}$$

$$\nu^i(n) = A_{\nu_1} cos\left[\omega_{\nu_1} n\right] \tag{5}$$

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

$$\begin{aligned}\Psi\left[\eta^i(n)\right] = {A^2}_{\eta_1} sen^2\left[\omega_{\eta_1}\right] + {A^2}_{\eta_2} sen^2\left[\omega_{\eta_2}\right] + \\ 2A_{\eta_1}A_{\eta_2}\Bigg(sen^2\left[\frac{\omega_{\eta_1}+\omega_{\eta_2}}{2}\right] cos\left[(\omega_{\eta_1}-\omega_{\eta_2})n\right] + \\ sen^2\left[\frac{\omega_{\eta_1}-\omega_{\eta_2}}{2}\right] cos\left[\left(\omega_{\eta_1}+\omega_{\eta_2}\right)n\right]\Bigg)\end{aligned} \tag{6}$$

$$\Psi\left[\nu^i(n)\right] = {A^2}_{\nu_1} sen^2\left[\omega_{\nu_1}\right] \tag{7}$$

Observando-se (6) e (7), percebe-se que o sinal resultante da aplicação do operador TEO no sinal voz sob estresse ($\nu^i(n)$) é uma constante, enquanto o resultante do sinal em estado Neutro ($\eta^i(n)$), é uma função do tempo consistindo de duas frequências $|\omega_{\eta_1} + \omega_{\eta_2}|$ e $|\omega_{\eta_1} - \omega_{\eta_2}|$. Portanto, a aplicação do operador TEO em bandas críticas pode detectar variações no padrão da frequência fundamental e formantes.

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 suavizar variações rápidas no sinal resultante após o operador TEO. Desta forma, mantem-se as informações sobre as variações na componente FM intactas, ou seja, mantendo-se informações sobre as frequências harmônicas e as suas interações.

A Fig. 3 ilustra o diagrama de blocos da extração do atributo TEO-CB-AUTO-ENV [11]. Um sinal de voz contendo apenas fonemas sonoros com alta energia é aplicado a um banco de filtros de Gabor baseados em informações sobre as bandas críticas, seguido da aplicação do operador TEO a cada um dos sinais filtrados. Cada sinal resultante da aplicação do operador TEO é então dividido em quadros de 25 ms, com sobreposição de 50%, e a área normalizada da autocorrelação é então calculada. A dimensão do vetor de atributos é igual ao número de filtros de Gabor utilizados.

Para aplicação em sistemas de RAL, a extração da Fig. 3 é dependente da informação fonética e, portanto, não aplicável. A Fig. 4 ilustra uma proposta de adaptação do atributo TEO-CB-AUTO-ENV, visando o acoplamento com a matriz de atributos contendo coeficientes MFCC. O sinal de voz é dividido em quadros utilizando-se a janela de Hamming, com duração de 20 ms e sobreposição de 50%. A extração é então realizada de forma análoga ao da Fig. 3, resultando em um vetor de atributos também com a mesma dimensão. Pode-se observar que a diferença entre as extrações se deve ao janelamento que é realizado antes da passagem pelo banco de filtros, ao tempo de duração do quadro e à remoção da restrição de aplicação apenas em sinais de voz sonoros.

III. Reconhecimento Automático de Locutor

Um sistema de RAL geralmente consiste das etapas de aquisição e pré-processamento do sinal de voz, extração de

TABELA I
NÚMERO DE TESTES DE 1S E DURAÇÃO MÉDIA (S) DA LOCUÇÃO DE TREINO UTILIZADOS NOS EXPERIMENTOS.

Emoções	Testes	Duração Média (s)
Raiva	124	15,4
Tédio	78	12
Medo	58	7,4
Neutro	62	10,3
Tristeza	77	12,1
Média	399	11,44

TABELA II
TAXA DE ACERTO DA IDENTIFICAÇÃO DE LOCUTOR COM AS EMOÇÕES PRIMARIAS(%) UTILIZANDO MFCC, TREINAMENTO COM LOCUÇÕES EM ESTADO NEUTRO E TESTES DE 1 S.

Teste	Modelo Neutro
Raiva	9,68
Tédio	67,95
Medo	20,96
Neutro	88,71
Tristeza	71,43
Média	51,75

atributos e classificação. Além disso é composto de duas fases: treinamento e teste. Na fase de treinamento são gerados os modelos dos locutores. Estes são utilizados na fase de teste para decidir sobre o reconhecimento da amostra em teste.

Na etapa de classificação, o sistema de RAL pode ser divido em duas tarefas: identificação e verificação. Na identificação, o sistema deve decidir qual o modelo mais provável dentre os cadastrados para a locução de teste, enquanto na tarefa de verificação, se a locução corresponde a identidade previamente declarada pelo locutor. Sistemas de RAL podem ainda ser clasificados como dependente ou independente do texto, dependendo da existência ou não de alguma restrição quanto ao tipo de locução que os usuários podem pronunciar [4]. Sistemas de RAL independentes do texto retratam situações mais próximas do caso real e, consequentemente, possuem um desempenho ligeiramente inferior ao caso dependente do texto e uma maior gama de aplicações.

Em ambientes sem a presença de ruído acústico, os sistemas de RAL, utilizando os coefientes MFCC e o classificador GMM, atingem taxas de acertos de identificação de até 99,5% [15]. Os coeficientes MFCC representam a energia do sinal de voz em bandas críticas igualmente espaçadas na escala Mel, que é baseada na resposta logarítmica da audição humana. A utilização do GMM como classificador em sistemas de RAL foi proposta em [14]. Cada componente gaussiana é capaz de representar formas espectrais capazes de modelar a identidade de um locutor.

Geralmente em experimentos de identificação de locutor a presença de estados emocionais nas locuções é negligenciada, pois estes são considerados informações paralinguísticas de alto nível. No entanto, tais efeitos estão sendo analisados em sistemas de reconhecimento de voz [16]. Baseando-se na observação da Fig. 2, é de se esperar que haja degradação na taxa de acerto.

IV. RESULTADOS

A. Base de voz

Para os experimentos de identificação de emoções e locutor, utilizou-se a base de dados pública *Berlin Emotional Database* (EMO-DB) [17]. 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.

As locuções foram subamostradas a 8 kHz e o silêncio foi removido. As locuções contendo sentenças mais longas foram utilizadas para o treinamento, enquanto as mais curtas foram segmentadas em locuções com 1 s de duração. Desta forma, os testes foram realizados em um cenário independente do texto. Na Tab. I, pode-se observar o número de testes de 1s e a duração média de cada locução utilizada no treinamento.

B. Identificação de Locutor

Para os testes de identificação de locutor, foram utilizados 16 coeficientes MFCC obtidos a partir de janelas de 20 ms com sobreposição de 50%. Os modelos GMM foram gerados com 16 distribuições gaussianas utilizando as locuções no estado Neutro (locução limpa).

A Tab. II apresenta os resultados das taxas de acertos de identificação de locutor. Pode-se perceber uma forte queda na taxa de acerto: de 88,71%, em seu estado Neutro, para 9,68% na presença da emoção Raiva. Os resultados comprovam que as alterações no trato vocal ocasionadas por estresse emocional afetam o reconhecimento automático de locutor. Portanto, a adição de atributos capazes de classificar o estresse na voz pode agregar informações ao RAL.

C. Identificação de Emoções

Para aplicação em um cenário de identificação de locutor, testes de identificação de emoções foram realizados num contexto dependente de locutor e multi-emoção. Ou seja, cada locutor possui um modelo para cada emoção, totalizando 50 modelos. A ordem do classificador de misturas gaussianas foi mantida inalterada (16). O critério de decisão adotado foi o da máxima verossimilhança, onde o modelo com a emoção de maior probabilidade a posteriori foi dada como saída do sistema. A extração de atributos utilizada foi a adaptação ilustrada na Fig. 4. Obteve-se um vetor de atributos de dimensão 16 por quadro.

A Tab. III mostra o resultado das taxas de acertos da identificação de emoções. Pode-se observar que o atributo consegue classificar as emoções de acordo com o eixo de ativação, mas não consegue distinguir com desempenho satisfatório emoções pertencentes ao mesmo eixo. Como exemplo, a taxa de acerto na identificação da emoção Medo (20,69%) é menor do que o erro de classificação para a emoção Raiva

TABELA III
TAXA DE ACERTO DE IDENTIFICAÇÃO DE EMOÇÕES PRIMÁRIAS(%) UTILIZANDO TEO-CB-AUTO-ENV PARA TESTES DE 1 S.

Testes	Raiva	Tédio	Medo	Neutro	Tristeza
Raiva	**82,26**	1,61	9,68	6,45	0
Tédio	10,26	**26,92**	15,38	25,64	21,79
Medo	39,66	15,52	**20,69**	15,52	8,62
Neutro	6,45	14,52	4,84	**43,55**	30,65
Tristeza	3,90	14,29	3,90	18,18	**59,74**

TABELA IV
TAXA DE ACERTO DA IDENTIFICAÇÃO DE LOCUTOR COM AS EMOÇÕES PRIMARIAS(%) UTILIZANDO MFCC+TEO-CB-AUTO-ENV, TREINAMENTO COM LOCUÇÕES EM ESTADO NEUTRO E TESTES DE 1 S.

Teste	Modelo Neutro
Raiva	**11,29**
Tédio	67,95
Medo	**24,14**
Neutro	**90,32**
Tristeza	**80,52**
Média	**54,84**

(39,66%). As maiores taxas de acerto de classificação foram das emoções Raiva, Tristeza e Neutro com 82,26%, 59,74% e 43,55%, respectivamente. Vale a pena ressaltar que foram utilizadas locuções de teste com duração de 1 s. A taxa média de acerto de classificação de emoções foi igual a 46,63%.

D. Identificação de emoções aplicada à identificação de locutor

Nesta seção, é analisado se o atributo TEO-CB-AUTO-ENV pode agregar informação à matriz de atributos MFCC. Vetores contendo 32 coeficientes (16 MFCC + 16 TEO-CB-AUTO-ENV) foram calculados e utilizados nos experimentos. Basicamente a matriz de atributos TEO-CB-AUTO-ENV foi acoplada a matriz MFCC. A ordem da mistura foi mantida. Os modelos gerados a partir de locuções com o estado Neutro foram utilizados, mantendo-se o mesmo cenário da identificação de locutor.

A Tab. IV apresenta os resultados das taxas de acertos da identificação de locutor para testes de 1 s. Pode-se perceber que a utilização do atributo baseada no operador TEO melhorou as taxas de acertos para todas as emoções, exceto Tédio que permaneceu inalterada. O maior aumento nas taxas de acertos ocorreu para a emoção Tristeza (9,09%). Pode-se então perceber que a utilização de atributos capazes de classificar emoções agregam informação ao RAL.

V. CONCLUSÕES

Neste trabalho foi demonstrado que a presença de estresse no sinal de voz, notadamente gerado por um estado emocional, pode diminuir o desempenho de sistemas de RAL no estado da arte. O descasamento entre as amostras de teste e treino geradas por emoções são tão degradantes quanto a presença de ruídos acústicos ambientais. Observou-se variações de até 79% na taxa de identificação de locutor na presença de emoções primárias nas locuções de teste.

Atributos baseados no operador TEO podem ser empregados na classificação emoções aplicado ao reconhecimento de locutor. Observou-se que o atributo pode discriminar emoções de acordo com o eixo de ativação. No entanto, o cenário multi-emoção ainda representa um desafio. Apesar do operador TEO ser melhor aplicado a sinais de voz contendo apenas sons sonoros, a sua adaptação para identificação de emoções em locuções curtas apresentou taxas de acerto de até 82% para a emoção Raiva. Para o caso multi-emoção, a taxa média de acerto de identificação de emoções foi igual a 46,11%.

Finalmente, a adição de atributos baseados no operador TEO agrega informação à matriz de coeficientes MFCC, tornando o sistema mais robusto a variabilidade do sinal de voz sob estresse. A aplicação do atributo TE0-CB-AUTO-ENV à identificação de locutor melhorou a taxa de reconhecimento em até 9%.

REFERÊNCIAS

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on circuits and systems for video technology*, January 2004.

[2] G. Doddington, "Speaker verification - identifying people by their voices," *Proceedings of the IEEE*, vol. 73, pp. 1651–1664, November 1985.

[3] J. Campbell, J.P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437 –1462, September 1997.

[4] 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, April 2004.

[5] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for enviroenmental robustness in speech recognition," *Speech Communications*, pp. 151–170, November 1996.

[6] E. Lombard, "Le signe de lelevation de la voix," *Ann maladie oreille larynx nez pharynx*, no. 37, pp. 101–119, 1911.

[7] J. C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon know as the lombard reflex," *Speech Communications*, pp. 13–22, November 1996.

[8] B. Yang and M. Lugger, "Emotional recognition from speech signals using new harmonic features," *Signal Processing*, vol. 90, pp. 1415–1423, 2010.

[9] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes and databases," *Pattern Recognition*, pp. 572–587, 2011.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, January 2001.

[11] G. Zhou, J. Hazen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 201–216, March 2001.

[12] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 599–601, October 2001.

[13] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'83)*, pp. 93–96, April 1983.

[14] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, pp. 72 –83, jan 1995.

[15] D. Reynolds, "Speaker identification and verification using gaussian mixture models," *Speech Communication*, no. 17, pp. 91–108, 1995.

[16] *Speech Communication, Special Issue on Speech Under Stress*, vol. 20, November 1996.

[17] 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.