Sistema Biométrico Multimodal Baseado em Análise Cepstral e Informação Espacial da Movimentação dos Lábios

Jefferson da Silva Costa[1], Verônica Nardy Paiva[1], Vitor Hugo Mendes Gomes[1], Adriano Vilela Barbosa[2], Glauco Ferreira Gazel Yared[1]

***Resumo*—Neste trabalho, propõe-se a implementação de um sistema biométrico multimodal baseado na combinação de informações acústicas parametrizadas por características conhecidas como coeficientes mel-cepstrais e informações visuais que representam a configuração labial durante a fala. Modelos individuais são estimados por um algoritmo de mistura de gaussianas, e a classificação é feita pelo método de maximização da expectativa. O desempenho e a robustez do sistema foram testados através da simulação de ambientes com diferentes níveis de ruído.**

Palavras-Chave*— *mel-cepstrais, audiovisual, biometria multi modal, GMM.

***Abstract*—This work proposes the implementation of a multi modal biometric system based on the combination of acoustic characteristics known as mel-cepstral coefficients and visual information representing the labial configuration during speech. Individual models are estimated with the gaussian mixture model algorithm, and the classification is obtained by the expectation maximization method. The overall performance is tested under different levels of corrupting noise.**

Key-Words*— *mel-cepstrals, audiovisual, multi modal biometry, GMM.

I. INTRODUÇÃO

A atual necessidade de se elevar a segurança no controle de acesso a sistemas de informação encontra no uso de características biométricas uma solução viável para a validação de usuários. Esse tipo de dado é facilmente coletado por meio de dispositivos como câmeras, microfones, sensores, etc..

No mais geral dos casos, em que as condições do ambiente em relação à iluminação e níveis de ruído, por exemplo, não podem ser garantidamente ideais, é possível observar uma queda no desempenho desses sistemas. Assim, a utilização de mais de uma característica biométrica pode ser usada para aumentar a robustez do sistema.

Neste trabalho propõe-se a implementação de um sistema biométrico multimodal baseado em informações audiovisuais do indivíduo, sendo extraídas das informações acústicas um conjunto de características conhecidas como coeficientes mel-cepstrais, além de informações espaciais extraídas dos sinais adquiridos a partir da movimentação facial durante a articulação da fala, analisando-se especialmente a região dos lábios. Propõe-se ainda a avaliação do desempenho geral do sistema com duas formas diferentes de associação dessas características, de maneira paralela, e de maneira concatenada, quando aplicadas a um algoritmo de classificação por modelo de mistura de gaussianas (GMM).

[1] Jefferson da Silva Costa, Verônica N. Paiva, Vitor Hugo M. Gomes, Glauco F. G. Yared UFOP - Campus João Monlevade, João Monlevade - Brasil. [2]Adriano V. Barbosa UFMG. Emails: jefferson.costa@aluno.ufop.edu.br, veronicanardy@hotmail.com, vitor.gomes@aluno.ufop.edu.br, adrianovilela@ufmg.br glauco@ufop.edu.br.

II. METODOLOGIA

A base dados audiovisual é composta por vinte indivíduos, sendo 10 de cada sexo. Para cada sujeito, tem-se os dados de vídeo e áudio da pronúncia de duas palavras do português brasileiro, “abrir” e “entrar”, com quinze repetições cada.

O sinal de áudio foi amostrado a uma frequência de 44,1 kHz, e o sinal de vídeo com 30 fps com resolução de 720x1280 pixels adquirido com iluminação natural. Então para cada usuário tem-se vetores com informações espaço-temporais de tamanhos diferentes, dependendo da velocidade da pronúncia em cada locução.

Os dados acústicos são parametrizados por meio dos coeficientes mel cepstrais (MFCC) [1], que devido a característica não linear da fala, são extraídos de trechos do sinal em janelas de 20ms, nos quais é possível assumir a linearidade [2], existindo entre janelas vizinhas uma sobreposição de 50% [4].

Com o auxílio do *Computer Vision toolbox* (*Matlab*) os lábios do indivíduo tem sua posição determinada e sua movimentação rastreada durante a produção da fala. Os vetores resultantes contém informações sobre a movimentação horizontal e vertical dos lábios. Em decorrência de variações da velocidade de locução entre os indivíduos, os sinais espaciais foram interpolados de maneira que possuíssem 120 amostras por vetor.

Aos sinais visuais já interpolados, é aplicado um banco de filtros não uniformes de 0-15 Hz a fim de se analisar com diferentes resoluções espectrais, o comportamento dos sinais em diferentes faixas de frequência de 0-2Hz, 1-5Hz, 2-6Hz, 3-8Hz e 5-15Hz. As larguras de banda de cada faixa foram determinadas de maneira experimental.

Os conjuntos de características audiovisuais são criados de duas maneiras, concatenadas(1) e paralelas(2):

1. Os sinais acústicos e visuais são processados em

janelas de 20ms com sobreposição de 50%. Assim, aos vetores de coeficientes mel-cepstrais de cada janela, são concatenadas características correspondentes ao movimento dos lábios adquirido simultaneamente, no mesmo intervalo de 20 ms resultando em um único vetor de características

2. A etapa de extração de características é feita de maneira separada para os sinais acústicos e visuais. Para os sinais acústicos, a extração dos coeficientes mel-cepstrais ainda é feita em janelas de 20ms com 50% de sobreposição, porém, esse janelamento não é mais aplicado aos sinais visuais.

Os modelos individuais são estimados pelo algoritmo de mistura de gaussianas, e a classificação é baseada no método de maximização da expectativa [3]. A robustez e capacidade de generalização dos modelos são avaliadas pela técnica *k-fold cross validation* com 10 repetições. Para a simulação de ambientes não ideais, são adicionados aos sinais acústicos diferentes níveis de ruído branco gaussiano.

III. RESULTADOS

Na análise comparativa entre os desempenhos do sistema com características puramente acústicas e o sistema audiovisual com características concatenadas, Figura 1, é possível observar um desempenho semelhante em condições de relação sinal ruído acima de 20 dB, onde as taxas de acerto obtidas são da ordem de 85% em média para ambos os modelos, atingindo uma taxa máxima de 92% para o sistema puramente acústico e 88% para o sistema audiovisual.

Em casos onde percebe-se baixa qualidade nos sinais acústicos (SNR abaixo de 0 dB), tem-se um desempenho insatisfatório de 8% em média para o sistema puramente acústico, atingindo uma taxa máxima de 12%, e para o sistema audiovisual um desempenho médio de 28%, atingindo um máximo de 36%, como visto na Figura 1.

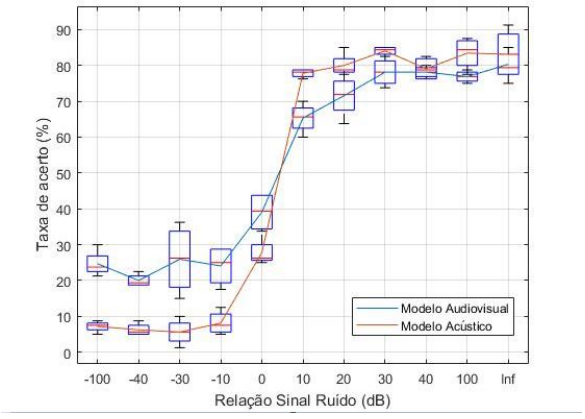

Figura 1: Desempenho do sistema audiovisual com características concatenadas

Na Figura 2, onde são comparados os desempenhos dos sistemas puramente acústico e audiovisual com características paralelas, nota-se que para valores de relação sinal ruído mais altos, o desempenho dos sistemas é bem semelhante, atingindo em ambos os casos aproximadamente de 80% em média. Porém, para níveis mais altos de ruído sonoro, nota-se uma queda considerável no desempenho do sistema acústico que chega a apresentar valores abaixo de 10%. O mesmo efeito não é visto no sistema audiovisual, que apresenta um desempenho mais estável, por volta de 80%, e independente do efeito de ruído sonoro aditivo.

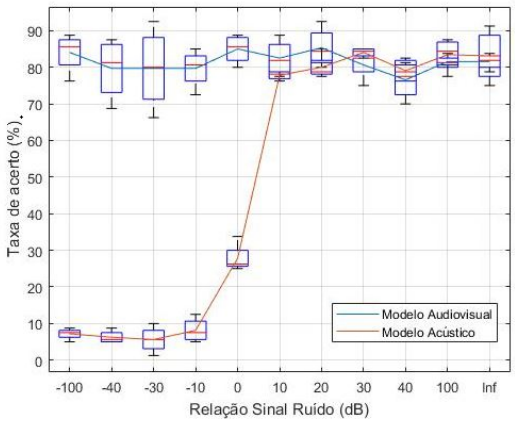

Figura 2: Desempenho do sistema audiovisual com características paralelas

IV. CONCLUSÕES

A análise dos sinais visuais em intervalos de 20ms simultâneos à locução não se mostra vantajosa para a implementação de um sistema biométrico multimodal, apesar de se observar uma melhoria de cerca de 16% (de 12 para 28%) em condições de baixa relação sinal-ruído. Por outro lado, na abordagem com a parametrização paralela dos sinais acústicos e visuais, observa-se que o desempenho do sistema audiovisual se mantém satisfatório, em torno de 80% em termos de taxa de acerto, mesmo diante de condições de baixa relação sinal-ruído, tornando o sistema mais robusto e, portanto, viável em termos práticos. Assim, a utilização de informações visuais da movimentação dos lábios durante a locução pode aumentar consideravelmente a robustez em sistemas biométricos de reconhecimento de locutor em ambientes ruidosos.

REFERÊNCIAS

[1] FUKADA, Toshiaki et al. An adaptive algorithm for mel-cepstral analysis of speech. In: Proc. ICASSP. 1992. p. 137-140.

[2] QUATIERI, Thomas F. *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[3] BISHOP, Christopher M. *Pattern recognition and machine learning*. springer, 2006.

[4] Frischholz, Robert W., and Ulrich Dieckmann. "BiolD: a multimodal biometric identification system." *Computer* 33.2 (2000): 64-68.

[5] GOPI, E. S. Digital speech processing using Matlab. New Delhi: Springer India, 2014.