Utilização da técnica MFCC em conjunto com os parâmetros extraídos do sinal glotal para melhorar o desempenho de um sistema de verificação de locutor

Filipe Moreira da Silveira, Carla Florentino Schueler e Edson Cataldo

***Resumo*—Esse artigo apresenta um algoritmo de verificação de locutor dependente de texto, que alia coeficientes MFC (*mel frequency cepstrum*) extraídos de sinais de voz a características extraídas do sinal glotal, utilizando modelos ocultos de Markov (*Hidden Marcov Models-HMM*) para identificar o padrão estocáticos dessas medidas. O sinal glotal, gerado imediatamente após a passagem do ar pelas cordas vocais, e é obtido nesse trabalho por filtragem inversa.**

Palavras-Chave*—*Verificação de locutor, MFCC, sinal glotal, HMM.

***Abstract*—This article presents the development of a text dependent speaker verification algorithm, which alias the Mel Frequency Cepstrum Coeficients with parameters calculated from the glottal signal using Hidden Markov Models to identify the stochastic pattern of these measurements. The glottal signal is generated when the air flow passes through the vocal chord, and is obtained in this work by inverse filtering.**

Keywords*—*Speaker verification, MFCC, Glottal signal, HMM.

I. INTRODUÇÃO

A voz é um fenômeno físico resultante da propagação do fluxo de ar, proveniente dos pulmões, através da glote e do trato vocal, com posterior irradiação pela boca [1]. A unicidade dos órgãos envolvidos no processo de geração da voz é o que torna a fala um sinal biométrico, podendo ser usada para a verificação de um locutor. Sistemas de verificação de locutor podem ser então construídos para os mais diversos fins, como por exemplo, para sistemas que utilizam a voz como chave de acesso a ambientes restritos. O algoritmo de verificação de locutor dependente de texto apresentado neste trabalho alia coeficientes MFC (*Mel Frequency Cepstrum*) extraídos de sinais de voz a características extraídas do sinal glotal, utilizando modelos ocultos de Markov (*Hidden Marcov Models-HMM*) para identificar o padrão estocásticos dessas medidas. O sinal glotal é a onda acústica gerada imediatamente após a passagem do ar pelas cordas vocais, e entende-se que possua informações características do locutor, pois ainda não sofreu a influência do trato vocal, possibilitando assim o aumento da eficiência do sistema [2]. Neste trabalho, este sinal é obtido minimizando-se os efeitos do trato vocal no sinal de voz por filtragem inversa, considerando-se a hipótese de que o trato vocal pode ser satisfatoriamente aproximado por um sistema linear e invariante no tempo para elocuções suficientemente pequenas.

Filipe Silveira, Bolsista do Programa Institucional de Bolsas de Iniciação à Inovação – PIBInova/PDI/UFF da Agência de Inovação/PROPPI, Carla Schueler e Edson Cataldo¸ Escola de Engenharia, Universidade Federal Fluminense (UFF), Niterói-RJ, Brasil, E-mails: carla_schueler@id.uff.br, filipesilveira@id.uff.br, ecataldo@im.uff.br .

II. OBJETIVO

Determinada uma palavra-senha, pretende-se que o sistema possa caracterizar a voz de um locutor específico, chamado aqui de locutor de interesse (Li), a partir de elocuções gravadas previamente, para então julgar se novos áudios são também amostras deste locutor. Além disso, observa-se que o desempenho do sistema aumente com o uso dos parâmetros do sinal glotal em conjunto com os MFCCs quando comparado com um sistema que use apenas MFCCs.

III. METODOLOGIA

A. Obtenção dos parâmetros da voz

Os MFCCs são coeficientes calculados a partir do espectro de frequências de pequenas janelas do sinal de voz, obtido por meio da FFT (*Fast Fourier Transform)* do sinal. Esse espectro de frequências é submetido a um banco de filtros triangulares, igualmente espaçados na escala de frequências Mel, para aproximar a percepção do sistema auditivo humano às frequências do som. Os coeficientes são calculados a partir da transformada discreta de cosseno aplicada à saída dos filtros.

Para a obtenção dos parâmetros do sinal glotal é preciso primeiro estima-lo a partir do sinal de voz. Nesta abordagem foi utilizado o algoritmo IAIF (*Iterative Adaptive Inverse Filtering*), que é composto de três blocos principais: a estimação dos efeitos do trato vocal por análise LPC (*Linear Predictive Coding*), a retirada desses efeitos do sinal por filtragem inversa e a eliminação dos efeitos da irradiação dos lábios por integração. Feita a recuperação do sinal glotal, os parâmetros são calculados usando a derivada e o espectro de frequência do sinal glotal.

B. Sistema de verificação de locutor

O sistema permite que um conjunto de amostras de áudio sejam gravadas e, após um pré-processamento, os MFCCs e parâmetros glotais são calculados para todo esse conjunto e utilizados para treinar um HMM que caracteriza o locutor. A implementação dos modelos ocultos de Markov foi atravez do *Hidden Markov Model (HMM) Toolbox for Matlab* [3].

A partir do modelo gerado podem ser calculadas verossimilhanças entre o modelo treinado e novas amostras de áudio. Essas medidas serão utilizadas para a construção de funções densidade de probabilidade (fdp) que serão usadas para definir os critérios de decisão avaliados pelo sistema [4]. A mais importante dessas fdps, gerada a partir de um grupo de

áudios exclusivamente do locutor de interesse, é a chamada curva padrão de equação y = *F(x)*. São usados dois critérios de avaliação: distância L_1, utilizada pelos desenvolvedores do sistema como um parâmetro para avaliarem sua funcionalidade, e o critério do limiar, utilizado posteriormente pelo programa pra decidir quando um locutor deve ou não ser aceito.

C. *Critérios de avaliação*

O primeiro critério, da distância L_1, realiza a verificação para um grupo de áudios teste do locutor candidato, do qual calculam-se as medidas de verossimilhança. Uma fdp dada por y = *G*(x) é calculada para essas medidas, e em seguida obtém-se a distância entre esta curva e a curva padrão $y = F(x)$ ao longo de todo o espaço *I* onde estão definidas. O cálculo da distância L_1 é dada pela Eq. (1).

Cálculo da distância L_1:

$$L_1 = \int_I |F(x) - G(x)| dx \qquad (1)$$

O segundo critério, chamado aqui de limiar, utiliza uma fdp $y = V(x)$, calculada a partir das verossimilhanças de um grupo de áudios de locutores mistos pré-armazenado. Adota-se como limiar de aceitação a abscissa do ponto de interseção entre esta fdp e a curva padrão, conforme ilustrado na Fig.1. Dessa forma, se um áudio teste possui uma verossimilhança maior que o limiar, então o mesmo é aceito como pertencente ao locutor de interesse, caso contrário ele é recusado. Esse critério é, então, usado para realizar a verificação segundo um único áudio, e não mais para um grupo deles como no caso da distância L_1.

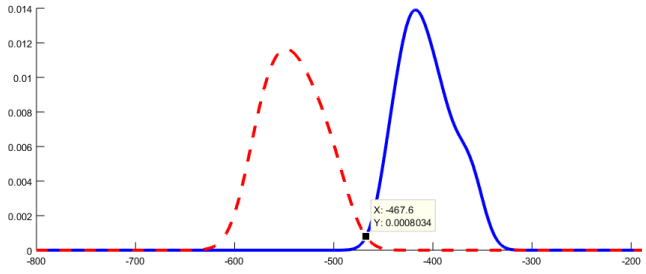

Fig. 1. Determinação do limiar pela interseção entre as fdps.

IV. RESULTADOS

Escolhida a sequência de vogais "aui" como palavra-chave, criou-se uma base de 200 elocuções do locutor de interesse. 60% dessa base foi usada para o treinamento do HMM, 20% utilizada para gerar a curva padrão e o restante para testar a acurácia do sistema. Para verificar a robustez do sistema foram também recolhidas amostras de áudios de outros três locutores desconhecidos.

O locutor desconhecido 1 (Ld_1) não possui parentesco com o locutor de interesse, além de ter sexo e idade distintos. O locutor desconhecido 2 (Ld_2) também não possui grau de parentesco, mas é do mesmo sexo e faixa etária. Já o locutor desconhecido 3 (Ld3) tem grau de parentesco próximo, mesmo sexo e faixa etária do locutor de interesse, tendo assim maior chance de gerar erros de interpretação pelo sistema.

Os resultados foram separados em dois grupos de testes, o primeiro com o modelo que utiliza apenas MFCCs e o segundo, que utiliza vetores híbridos de MFCCs e parâmetros obtidos do sinal glotal.

TABELA I. RESULTADOS SEGUNDO O CRITÉRIO DA DISTÂNCIA L_1

Locutores	Sem parâmetros glotais	Com parâmetros glotais
Li	0.4328	0.2540
Ld_1	1.9308	1.9703
Ld_2	1.9507	1.9964
Ld_3	1.8875	1.8898

O valor do limiar calculado para o modelo que usa exclusivamente MFCCs foi -318, enquanto o limiar encontrado com o uso das duas técnicas em conjunto foi -476. O percentual de acerto do sistema pode ser visto na Tabela II.

TABELA II. RESULTADOS SEGUNDO O CRITÉRIO DO LIMIAR

Locutores	Percentual de acerto	
	Sem parâmetros glotais	Com parâmetros glotais
Li	100%	87%
Ld_1	100%	100%
Ld_2	100%	100%
Ld_3	85%	100%

V. CONCLUSÕES

Os resultados segundo a distância L_1 mostram que as curvas dos locutores desconhecidos são mais distantes da curva padrão quando utilizada a técnica MFCC em conjunto com os parâmetros do sinal glotal do que quando apenas a técnica MFCC é utilizada. Esse resultado permite que o algoritmo de verificação de locutor aceite mais vezes os áudios do locutor de interesse e recuse completamente os áudios dos locutores desconhecidos pelo modelo, caracterizando a robustez do sistema.

Para o critério do limiar, os resultados com apenas a técnica MFCC apresentaram alguns aceites de áudios do locutor desconhecido 3. Esse erro foi mitigado com a inclusão dos parâmetros glotais. É possível perceber, no entanto, que ao utilizar as técnicas conjuntamente, alguns áudios do locutor de interesse passaram a ser rejeitados pelo sistema. Esse cenário é mais desejável do que o anterior, pois para sistema de segurança é preferível que o locutor de interesse precise repetir o teste algumas vezes para ser aceito, do que permitir o acesso de um locutor desconhecido.

Tal fato é inovador, pois apesar da ideia que os parâmetros extraídos do sinal glotal carregam mais informações sobre o locutor, pouco existe na literatura sobre a aplicação dessa técnica a fim de melhorar processos de verificação de locutor como o desenvolvido.

REFERÊNCIAS

[1] L. Rabiner e H. H. Wang, *Fundamentals of Speech Recognition*, Englewood Cliffs, N.J., Prentice Hall, 1994.

[2] L. Mendoza, E. Cataldo, M. Vellasco, M. Silva e J. Apolinario, "Classification of Vocal Aging Using Parameters Extracted From the Glottal Signal", *Journal of Voice*, v. 28, pp. 532–537, Maio 1999.

[3] K. Murphy, *Hidden Markov Model (HMM) Toolbox for Matlab*, http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html, 1988.

[4] E. Cataldo, C. Soize, and R. Sampaio., "A computational method for updating a probabilistic model of an uncertain parameter in a voice production model." *Uncertainties 2012*. Maresias, Brazil, p. 1-8, 2012.