ANÁLISE DE SISTEMAS DE RECONHECIMENTO AUTOMÁTICO DO LOCUTOR USANDO MFCC E QUANTIZAÇÃO VETORIAL

Mateus Lichfett Machado, Universidade Federal de Uberlândia, Laboratório de Acústica e Vibrações, Avenida João Naves de Avila nº 2160, 38400-089, Uberlândia-MG, mateus.lichfett@gmail.com

Marcus Antonio Viana Duarte, Universidade Federal de Uberlândia, Laboratório de Acústica e Vibrações, Avenida João Naves de Avila nº 2160, 38400-089, Uberlândia-MG, mvduarte@mecanica.com

Resumo. *Sistemas de Reconhecimento Automático de Locutor (RAL), embora complexos, são vastamente utilizados em processos de análise biométrica. A singularidade presente nas propriedades da voz de cada indivíduo decorre em função dos diferentes formatos entre os aparelhos de produção da fala humana (cavidade oral e nasal, traqueia, etc.) e da variação na quantidade de pressão utilizada para enunciação de palavras. Esta dessemelhança condiciona a voz como instrumento de identificação, garantindo sua utilização em serviços que requerem certo nível de segurança. A acuracidade destes sistemas (RAL) é imprescindível e, está mormente condicionada às técnicas empregadas para extração das características do sinal de voz e reconhecimento de padrões. O âmbito deste estudo compreende uma metodologia científica acerca do uso de duas técnicas: Coeficientes Mel-cepstrais (MFCC) e Quantização Vetorial (VQ) para extração das propriedades acústicas do sinal e reconhecimento de padrões, respectivamente. A fim de estudar a eficiência destes métodos, reproduziu-se um experimento com 8 pessoas, dentre as quais apenas 4 tiveram suas vozes catalogadas para ensaio. Os resultados obtidos comprovam a eficiência destas técnicas para identificação do locutor.*

Palavras chave*: Reconhecimento Automático de Locutor, Coeficientes Mel-cepstrais, Quantização Vetorial.*

1. INTRODUÇÃO

Vozes de pessoas diferentes também soam de maneiras distintas e, esta característica peculiar é o que permite que indivíduos sejam identificados apenas através de suas falas. A fala humana, assim como sua interpretação, é um fenômeno de elevada complexidade e importância, por isto importantes centros de tecnologia investem em pesquisas neste ramo.

Sistemas de Reconhecimento Automático de Locutor (RAL) apresentam elevada relevância em diversos campos que requerem biometria em função de seu elevado grau de confiança e baixo custo. Começaram a ser desenvolvidos há mais de 30 anos, e foram aprimorados com a evolução de processadores computacionais, com confiabilidade superior à 99%.

A identificação do sinal consiste em um processo de conversão da fala em forma de onda, em um sinal digitalizado, do qual são extraídos as características acústicas úteis para posterior processamento e análise de correlação com outro sinal para reconhecimento de padrões. A capacidade de capturar as relações de tempo, frequência e energia em um conjunto de coeficientes para análises Cepstrais é o que favorece o uso da técnica Coeficientes Mel-cepstrais (MFCC), pois permite extrair coeficientes correspondentes à faixas de frequência posicionados de forma logaritmica em uma escala chamada Mel, capaz de definir frequências de tons, próximos à percepção do sistema auditivo humano. Ademais, como técnica de congruência entre sinais correspondentes ao mesmo indivíduo, *Hidden Markov Models* (HMM), *Gaussian Mixture Models* (GMMs) ou Quantização Vetorial (VQ) estão sendo comumente adotadas (Huang, Acero e Hon, 2001).

O presente trabalho fora segmentado em três partes principais: Fundamentos Teóricos; Experimento e Análise de Resultados e Conclusões. Na primeira seção, será introduzido os princípios de reconhecimento do locutor através da voz, discutindo as nuances entre as fases de treinamento e teste; e ainda serão abordados fundamentos teóricos que governam as técnicas de extração das características de sinal (MFCC) e reconhecimento de padrões (VQ). Posteriormente, será apresentado o experimento, detalhando as condições de ensaio e, esboçando testes realizados. Finalmente, serão feitas observações ponderadas acerca dos resultados obtidos.

2. PRINCÍPIOS DO RECONHECIMENTO DE VOZ

Sistemas de Reconhecimento Automático de Locutor (RAL) operam para a identificação ou verificação imediata daquele que enuncia (Keshet e Bengio, 2009). Estes sistemas são fragmentados em dois módulos subsequentes: extração das propriedades acústicas do sinal e reconhecimento de padrões. Em uma primeira instância, dados importantes para quantificação do sinal de voz são adquiridos construindo-se um vetor acústico. A segunda etapa envolve o próprio procedimento de identificação do locutor, no qual distâncias entre o vetor acústico correspondente ao sinal de entrada e o conjunto de regiões, cada qual, associado a um locutor específico e construídas através de Quantização Vetorial são calculadas. A menor distância obtida, quando inferior à um certo limiar previamente definido, indicará quem é aquele que enuncia (Mafra. A, 2002).

2.1. Extração de Características de um Sinal de Voz

O primeiro passo em um sistema de Reconhecimento Automático de Locutor após a captação do sinal de voz é extrair as propriedades e identificar os componentes úteis do sinal de áudio para identificação linguística. Para melhor se adaptar às características da fala é preciso entender que sons produzidos por um ser humano são filtrados pelo aparelho vocal (constituído pelas cavidades orais, nasais, língua, traquéia, dentes etc.) e o formato deste é o que determina qual som é reproduzido. Modelando-o com precisão, obtém-se uma representação contundente ao fonema que está sendo reproduzido. Sob esta perspectiva, Coeficientes Mel-cepstrais (MFCC) são representações na qual a forma do trato vocal se manifesta no envelope do espectro de potência, em um curto espaço de tempo. (DHINGRA, S. et. Al, 2013).

Para implementação dos Coeficientes Mel-cepstrais é necessário a realização de uma sequência de passos. Como um sinal de áudio está constantemente variando é necessário estabelecer que em um curtíssimo período de tempo, ele seja quase estacionário. Discretiza-se o sinal, enquadrando-o em frames, comumente variantes de 20 à 40 ms. Neste trabalho, optou-se por enquadrar o sinal em frames de 30 ms. Cada frame contém 256 amostras sobrepostos à cada 100 amostras.

O próximo passo, consiste em realizar um janelamento de cada frame, minimizando as descontinuidades, ou distorções espectrais do sinal tanto no início quanto no fim de cada frame, afunilando-o para zero nestes limites. Definindo a janela como $w(n),\ 0,\leq n \leq N-1$, com N sendo o número de amostras em cada frame, o resultado do janelamento é o sinal:

$$y_i(n) = x_i(n) \cdot w(n), \qquad 0,\leq n \leq N-1 \tag{1}$$

O tipo de janela selecionado fora *Hamming*, definida como:

$$w(n) = 0{,}54 - 0{,}56\cos\left(\frac{2\pi\pi}{N-1}\right), \qquad 0,\leq n \leq N-1 \tag{2}$$

Após o janelamento, calcula-se o espectro de potência de cada frame em um periodograma que identifica quais frequências estão presentes naquele sinal, operando similarmente à cóclea humana. Sequencialmente, converte-se cada frame do domínio do tempo para o da frequência, utilizando uma Transformada Rápida de Fourier (FFT), definida sobre o conjunto de N amostras em cada frame, (x_n) segundo a Eq. (3):

$$xk = \sum_{n=0}^{N-1} x_n \cdot e^{-j\,2\pi\pi k/N}, \qquad k = 0,1,2,\ldots,N-1 \tag{3}$$

Considera-se os valores absolutos dos resultados obtidos (magnitudes das frequências), uma vez que X_k contém valores complexos. O resultado obtido é chamado de periodograma espectral e suas estimativas ainda contêm informações inúteis para sistemas *RAL*, sendo necessário filtrá-las. Além disto, o periodogama é incapaz de discernir frequências estreitamente espaçadas. Assim, agrega-se conjuntos de frequências próximas para calcular a quantidade de energia contida em certas regiões. Este cálculo é feito através de bancos de filtros Mel, ilustrado aqui pela Fig. (1). Tanto na Figura, quanto no trabalho, utiliza-se 20 filtros. O primeiro filtro é bastante estreito, sua função está em calcular a energia contida em regiões de pequenas frequências, próximas à 0 Hz. Conforme as frequências aumentam, os filtros subsequentes tornam-se mais largos, uma vez que não há uma grande importância com as frequências mais elevadas quando se estuda um sinal de voz.

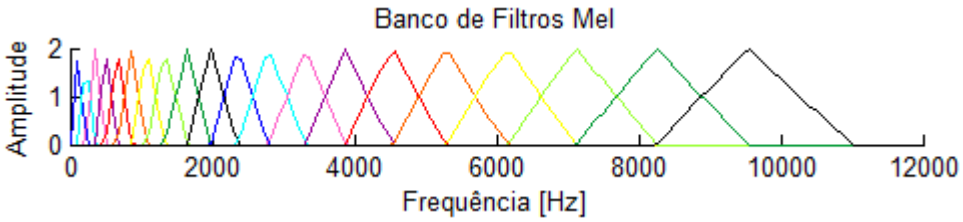

Figura 1. Ilustração dos bancos de filtro Mel para 20 coeficientes espectrais de potência.

Como espaçar os filtros adjacentes, e quão largos estes serão é solucionado usando a escala de Mel, Eq. (4).

$$M(f) = 1125 \cdot \ln(1 + f/700) \tag{4}$$

Adquiridos as energias dos bancos de filtros, aplica-se a eles uma função logarítmica, para normalizar os resultados. Estudos psicofísicos apontam que a percepção humana de frequências contidas em um sinal de fala não segue um padrão linear em todo seu escopo. Até a frequência de 1000 Hz, o reconhecimento dos tons obedece uma certa linearidade, acima desta, apresenta um comportamento logarítmico. Estes fatores são reproduzidos pela escala Mel. (Gupta e Bansal, 2013).

O passo final é converter os dados até aqui obtidos de volta para o domínio do tempo aplicando uma Transformada Discreta do Cosseno (DCT). Se denotarmos os coeficientes de espectro de potência Mel, calculados no passo anterior

como $\tilde{S}_k$, $k = 0,2,\ldots,K-1$, em que *K* representa o número total de coeficientes (*K=20*), os coeficientes Mel-cepstrais (MFCC), denotados como $\tilde{c}_n$, podem ser calculados conforme a Eq. (5).

$$\tilde{c}_n = \sum_{k=1}^{K} log\left(\tilde{S}_k\right) \cdot cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], \qquad n = 0,1,2,\ldots,K-1 \tag{5}$$

Exclui-se o primeiro coeficiente do espectro de potência do cálculo para Transformada Discreta do Cosseno, pois representa o valor médio do sinal de entrada, que contém pouca informação sobre o Locutor. A *DCT* dos coeficientes espectrais de potência é calculada por duas razões principais. Primeiro, por razão do *overlap* entre os frames que se utilizou para discretizar o sinal, as energias dos bancos de filtros calculadas estão correlacionadas em certo grau, e aplicando a *DCT*, as energias tornam-se não correlacionadas, possibilitando a modelagem das características do sinal em um sistema de reconhecimento de padrões. E segundo, como maneira de melhorar a eficiência do sistema, uma vez que apenas 12 dos 26 coeficientes obtidos pela *DCT* são mantidos. Coeficientes mais elevados da *DCT* representam rápidas mudanças nas energias dos bancos de filtros, degradando a performance de sistemas *RAL*.

2.2. Reconhecimento de Padrões

Esta etapa consiste em identificar o locutor através do reconhecimento de padrões entre as características extraídas do sinal de entrada e daquelas previamente adquiridas e armazenadas em uma biblioteca digital. Este método classifica objetos de interesse em um certo número de categorias, ou classes. Os objetos de interesse são chamados de padrões, que são, na verdade, sequencias vetoriais que contém as propriedades acústicas previamente extraídas (vetores acústicos), conforme descrito na subseção anterior. As classes, por sua vez, referem-se aos locutores (ARYA, S. et. Al, 1987).

A escolha da técnica utilizada para reconhecimento de padrões neste estudo fora Quantização Vetorial (*VQ*), por apresentar rápida compilação, operando com dados comprimidos, além de fácil implementação e alto índice de acertos (Makhoul, J. et. Al, 1985). A técnica permite a modelagem de funções de densidade probabilística através da distribuição de vetores de classificação. É um processo de mapeamento de um espaço vetorial para pequenas regiões finitas que constam naquele espaço, governadas por um centróide, utilizando vetores acústicos obtidos dos sinais de voz dos indivíduos treinados. Cada região é chamada de *cluster* e seu centróide chamado de *codeword*. Uma coleção de *codewords* forma um *codebook*. O método utilizado para construção de *VQ codebooks* é o algoritmo *LBG* (LINDE; BUZO; GRAY, 1980), que opera agrupando um conjunto de *L* vetores de treinamentos em um conjunto de *M* vetores *codebook*.

A implementação do algoritmo *LBG* projeta intuitivamente um *codebook* de *M* vetores em estágios subsequentes. Isto ocorre de maneira recursiva: primeiramente, designa-se um vetor *codebook*, que será o centróide de todo o conjunto de treinamentos; em seguida, dobra-se o tamanho do *codebook*, repartindo cada *codeword* $\left(y_n\right)$ segundo as Eq. (6).

$$y_n^+ = y_n\left(1+\varepsilon\right) \tag{6.a}$$

$$y_n^- = y_n\left(1-\varepsilon\right) \tag{6.b}$$

Em que, *n* varia de 1 ao tamanho do *codebook*, e ε é um parâmetro de repartição (normalmente, $\varepsilon = 0{,}01$). O terceiro passo está em encontrar o vizinho mais próximo, ou seja, para cada vetor de treinamento, procura-se a *codeword* no presente *codebook* mais próximo e assimila-se o vetor à correspondente célula. O quarto passo, refere-se à atualização do centróide em cada célula, utilizando o centróide dos vetores de treinamento atribuídos àquela célula. Repete-se então, os passos 3 e 4 até que a distância média caia abaixo de um limiar anteriormente definido (*k=5*). E, finalmente, repete-se os passos 2, 3 e 4 até que se atinja uma matriz de *codebooks* com *M* vetores.

Na fase de reconhecimento de locutor, um sinal de entrada de uma locução desconhecida é tratada com VQ, usando cada *codebook* de treinamento. Calcula-se então a distância do vetor correspondente ao sinal desconhecido às *codewords* (distância esta chamada de distorção VQ). E, assim, o indivíduo que corresponde ao *codebook* com a menor distorção VQ é identificado como aquele que enunciou o sinal de entrada (GERSHO; GRAY, 1992).

3. EXPERIMENTO E ANÁLISE DE RESULTADOS

3.1. Descrição do Experimento

Para avaliar a eficiência da aplicação dos métodos em estudo que identifica automaticamente o locutor, desenvolveu-se um programa iterativo na plataforma Matlab v.7.12.0 (R2011a). O programa consta de três funções principais, permitindo ao usuário: treinar sua voz e integrar o sinal processado ao banco de dados; enunciar comandos para automática identificação do locutor; e analisar dados técnicos, como detalhes da densidade espectral de cada sinal, ou mesmo, ouvir qualquer sinal que esteja disposto na biblioteca digital.

O experimento realizado para avaliar a eficácia dos métodos contou com a participação de 8 pessoas voluntárias. Deste grupo, 4 pessoas tiveram sua voz treinadas (indivíduos 1 à 4), enunciando frases com duração de 5 segundos. Para

adquirir o sinal de entrada, utilizou-se o microfone embutido ao notebook ASUS modelo X53sv (mono); uma amostragem de frequências de 22,05 kHz; e 8 bits por amostra. Todos os testes foram conduzidos em ambientes controlados. A Fig. (3) ilustra o sinal de voz capturado pelo indivíduo 1, os dados utilizados para amostragem do sinal, bem como seu processamento, com a seguinte locução: "Seu sonho é tão palpável, quão grande sua vontade de realizá-lo".

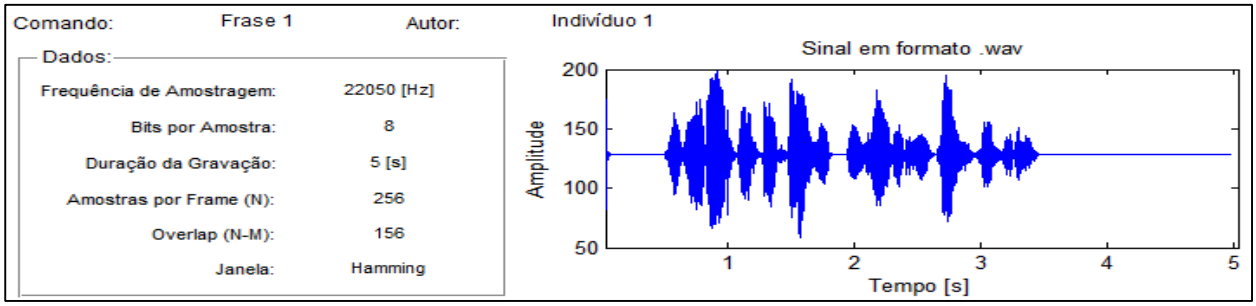

Figura 3. Ilustração do sinal de voz enunciado pelo indivíduo 1.

Após a captação dos sinais, extração das propriedades acústicas e montagem do *codebook* para cada indivíduo registrado, iniciou-se a fase de testes. Para tal, orientou-se, em uma primeira instância que cada indivíduo enunciasse livremente frases quaisquer e, posteriormente, aos indivíduos 1 à 4, que repetissem a mesma frase utilizada para a fase de treinamento. A Tabela (1) apresenta os comandos utilizados pelos indivíduos cujas vozes foram treinadas.

Tabela 1. Frases enunciadas pelos indivíduos cujas vozes foram treinadas.

Indivíduos Treinados	Indivíduos	Frases	Duração
	1	(1) "Seu sonho é tão palpável, quão grande sua vontade de realizá-lo."	5s
	2	(2) "Tudo que você precisa é de amor."	5s
	3	(3) "Felicidade, só é real, quando compartilhada."	5s
	4	(4) "Poesia, beleza, romance e amor. Por estas razões, continuamos vivendo."	5s

3.2. Resultados e Comentários

Para auferir o resultado do experimento em questão, realizou-se ao total 53 testes, dos quais, cada participante enunciou comandos ao menos 5 vezes. Em todos os testes, os resultados corresponderam à correta avaliação do sistema de reconhecimento de locutor. Isto é, quando um locutor reconhecido pelo sistema (que tivera sua voz treinada) enuncia algum comando, sua voz é detectada e associada corretamente ao sinal de treinamento. E, quando algum indivíduo desconhecido para o sistema, da mesma maneira, enuncia um comando, o sistema envia uma mensagem, alertando que aquele locutor era desconhecido. Na Tab. (2), fora apresentado alguns resultados obtidos. Na coluna de frases, as frases numeradas estão indicadas na Tab. (1), e correspondem respectivamente aos indivíduos com mesma numeração, enquanto 'Qualquer' designa que o locutor não enunciou nenhuma das frases catalogadas.

Tabela 2. Distorção VQ calculadas para comandos de voz de entrada.

Locutor (Ind.)	Frases	Distância do sinal de voz às *codewords* (distorção VQ)				Acertos
		Ind. 1	Ind. 2	Ind. 3	Ind. 4	
1	Frase (1)	2,103	6,445	7,689	10,511	✓
2	Frase (2)	5,311	2,243	8.874	6,459	✓
2	Qualquer	6,886	3,547	7,552	7,105	✓
3	Frase (3)	6,103	5,744	2,678	9,201	✓
3	Frase (4)	7,221	6,781	3,451	7,420	✓
5	Frase (1)	9,311	8,219	6,443	8,197	✓
6	Frase (1)	7,581	6,172	10,170	7,111	✓
7	Frase (2)	6,850	6,783	7,866	6,451	✓
6	Frase (3)	7,002	6,872	8,193	6,339	✓
4	Frase (1)	6,991	5,446	6,204	3,201	✓
8	Qualquer	7,455	7,233	6,507	6,893	✓

Optou-se por apresentar os resultados esboçados na Tab. (2) com o propósito de ilustrar algumas distâncias calculadas em relação aos *codebooks* gerados para cada indivíduo que tivera sua voz treinada. A menor distância calculada designaria o locutor quando esta distância fosse inferior ao limiar proposto ($k=4$). Além disto, observa-se que, quando indivíduos que tiveram suas vozes treinadas enunciam o mesmo comando que quando utilizaram para treinar, a distância é inferior às distâncias calculadas em momentos que o locutor enuncia outro comando que não aquele que fora treinado.

4. CONCLUSÕES

O máximo índice de acertos para este experimento comprova a eficácia da combinação das técnicas Coeficientes Mel-cepstrais (MFCC) e Quantização Vetorial (VQ) para sistemas de reconhecimento automático de locutor sob condições de ruído controlado. Observou-se também, que o método aumenta sua capacidade de identificação quando o locutor que fora cadastrado no sistema enuncia o mesmo comando e, de maneira semelhante ao que ele utilizara para treinamento, isto é, quando ele enuncia comandos sem prolongação de sílabas e sem alterar a maneira de falar. Sugere-se aos trabalhos futuros, o emprego de técnicas como Dynamic Time Warping (DTW) para que condições de diferenças de tempo ao enunciar as sílabas não seja um fator de impedimento para máxima capacidade de identificação do sistema.

5. REFERÊNCIAS

ARYA, S.; MOUNT, D. M. Algorithms for Fast Vector Quantization. Proc. Data Compression Conference, J. A. Storer and M. Cohn, eds. Snowbird, Utah, 1993, IEEE Computer Society Press, pp 381-390.

DHINGRA, S. D.; NIJHAWAN, G.; PANDIT, P. Isolated Speech Recognition Using MFCC and DTW. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.2, Issue 8, 2013.

GERSHO, A.; GRAY, R. M. Vector Quantization and Signal Compression. Editora Springer-US, 1992.

GUPTA S.; BANSAL A. Et al. Feature Extraction Using MFCC. Signal & Image Processing: An International Journal (SIPIJ) Vol.4, No.4, Agosto 2013.

HUANG, X.; ACERO, A.; HON, H.-W. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Editora Prentice Hall PTR, 1ª Ed. 2001.

KESHET, J.; BENGIO, S. Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods. Editora John Wiley and Sons, 1ª Ed. 2009.

MAFRA, A. T. Reconhecimento Automático de Locutor em Modo Independente de Texto Por Self-organizing Maps. Dissertação em Engenharia Mecânica pela Escola Politécnica da Universidade de São Paulo. São Paulo-SP. 2002.

MAKHOUL, J.; ROUCOS, S.; GISH, H. Vector Quantization in Speech Coding. Proceedings of the IEEE, Vol.73, No.11, pp 1551-1588, Novembro 1985.

6. AGRADECIMENTOS

Os autores agradecem à Universidade Federal de Uberlândia, Faculdade de Engenharia Mecânica, Laboratório de Acústica e Vibrações (LAV) pela infra-estrutura provida; à CAPES pelo apoio financeiro e aos voluntários que emprestaram suas vozes para a realização do experimento.

7. ABSTRACT

Automatic Speaker Recognition Systems (ASR), though complex, are widely used in biometric analysis processes. The uniqueness of each individual speech characteristics occurs depending on the different shapes of the human speech production apparatus (nasal and oral cavities, trachea, etc.), and the variation in the amount of pressure used for uttering words. This dissimilarity allows the use of voice as an identification tool, ensuring their use in services that require some level of security. The accuracy of these systems (ASR) is essential for its application, and is mainly conditioned by the techniques used for feature extraction of the speech signal, and pattern recognition. The scope of this study comprises a scientific methodology on the use of two of these techniques: Mel-cepstral coefficients (MFCC), and Vector Quantization (VQ), for signal feature extraction and pattern recognition respectively. In order to study the efficiency of this method, an experiment was reproduced carrying out with 8 people, of which only 4 had their voices cataloged for experiment. The results obtained prove the efficiency of these techniques for ASR systems.

***Keywords**: Automatic Speaker Recognition, Mel-cepstrum Coefficients, Vector Quantization.*

8. RESPONSABILIDADE PELAS INFORMAÇÕES

Os autores são os únicos responsáveis pelas informações incluídas neste trabalho.