

Reconhecimento de Locutor pela Voz usando o Classificador Polinomial e Quantização Vetorial

Wemerson D. Parreira*

Faculdade de Engenharia Elétrica, UFU,
38400-902, Uberlândia, MG

E-mail: wdparreira@yahoo.com.br, wemerson@rexlab.ufsc.br,

Gilberto A. Carrijo

Faculdade de Engenharia Elétrica, UFU
Av. João Naves de Ávila, 2121
38400-902, Uberlândia, MG

E-mail: gilberto@ufu.br.

Resumo

O Reconhecimento de locutores pela voz é um problema de alta complexidade. Um método que utiliza classificadores polinomiais e quantização vetorial foi desenvolvido neste trabalho para a solução deste problema. A aplicação inicial da quantização vetorial permite a não necessidade do alinhamento temporal das amostras dos sinais de voz, produzindo um conjunto de matrizes que serão as entradas do sistema baseado em classificadores polinomiais. Os classificadores polinomiais têm excelentes propriedades classificatórias, devido ao Teorema de Weierstrass, eles são aproximações universais dos classificadores ótimos de Bayes. Será mostrado que a combinação destes dois métodos obteve um rendimento considerável em relação à quantização vetorial simples no reconhecimento de locutores.

1 Introdução

É conhecido que vozes de pessoas diferentes soam de maneira também diferentes. Esta importante propriedade faz com que se possa distinguir uma pessoa da outra apenas pela sua voz.

A técnica de reconhecer uma pessoa pela sua voz é conhecida como reconhecimento automático do locutor pela voz. O Reconhecimento Automático do Locutor, *Automatic Speaker Recognition* (ASR), pela voz é uma técnica que teve início a mais de 30 anos, [1] e [2]. Mas recentemente com o desenvolvimento da integração em larga escala e com processadores de alta velocidade foi possível implementar as técnicas teóricas até então desenvolvidas. Hoje vários sistemas estão sendo usados de maneira comercial onde a porcentagem de reconhecimento correto pode chegar até 99%, [3].

Em várias aplicações de reconhecimento da fala é muito difícil ter um desempenho que se aproxima do ser humano, mas no ASR o sucesso das máquinas é superior ao do homem. Hoje, no entanto várias pesquisas estão direcionadas a entender como uma pessoa distingue um locutor dentre vários outros. A maneira como se reconhece um locutor pode ser apresentada de duas maneiras:

Verificação Automática do Locutor (ASV)

Identificação Automática do Locutor (ASI)

A ASV é usada para verificar se a pessoa que reivindica o reconhecimento é realmente a pessoa, e não um impostor. Isto pode acontecer quando uma pessoa digita um código e logo em seguida fala uma frase. Ela tem por finalidade reconhecer se a voz é realmente da pessoa que é proprietária do código ou se é um impostor. Trata portanto de uma tarefa mais simples, pois compara um padrão teste com um padrão de referência envolvendo uma decisão binária, ou seja, a resposta será *sim* ou *não*.

O processo de ASI é usado quando se deseja reconhecer uma pessoa dentre um conjunto de várias outras pessoas, mas sem inicialmente fornecer qualquer informação, ou código da pessoa que se deseja identificar. Ela escolhe dentre um conjunto de N locutores qual deles o padrão em teste melhor se aproxima. Desde que N comparações são necessárias, a taxa de erro no sistema ASI pode ser mais alta do que no sistema ASV.

O reconhecimento do locutor pela voz pode ser feito através do uso de um texto conhecido (dependente do texto), ou pode ser feito através de um texto qualquer (independente do texto).

Um sinal da fala é produzido como resultado de uma seqüência complexa de transformações ocorrendo em diferentes níveis: semântica, lingüística,

*bolsista de Mestrado CNPq

articulação, e acústica. Em geral diferenças nestas transformações produzem diferenças nas propriedades do sinal da fala. Variações diferentes dos locutores são provenientes de diferentes cavidades vocais e diferentes hábitos das pessoas.

Além das variações entre os locutores (interlocutor) têm-se as variações com um mesmo locutor (intra-locutor) quando o mesmo pronuncia a mesma frase. Isto acontece dependendo do estado físico e emocional da pessoa.

Para reduzir as variações com um mesmo locutor é comum o uso do reconhecimento dependente do texto. A tarefa de verificação é realizada com a comparação de um texto falado no momento do reconhecimento com outro previamente gravado pelos locutores.

Este trabalho tem como objetivo utilizar o classificador polinomial, técnica já utilizada na literatura no reconhecimento de locutor, aplicado a coeficientes obtidos dos *codebooks* gerados no processo de quantização vetorial. Com o intuito de obter um maior rendimento aos apresentados por [6] e [11].

1.1 Técnicas Existentes

As técnicas para comparação de padrões mais conhecidas na literatura são: as estatísticas e as determinísticas. Nas técnicas estatísticas as comparações de padrões são feitas pela medida da função verossimilhança, ou probabilidade condicional, da observação do modelo. Nas técnicas determinísticas, o padrão é assumido ser uma réplica perfeita e o processo de alinhamento faz-se necessário para calcular a distância.

Os principais Métodos Determinísticos são os baseados em *Dynamic Time Warping* - DTW [5]; Quantização Vetorial - QV [6]; Redes Neurais [7] e Classificadores Polinomiais [8]. Dentre os estatísticos destacam-se os baseados em Função Densidade de Probabilidade e *Hidden Markov Models* - HMM [4].

1.2 O Reconhecimento

As técnicas de reconhecimento de locutor pela voz tiveram início a mais de 30 anos. Hoje com o desenvolvimento de microprocessadores muitas técnicas teóricas puderam ser implementadas.

O processo de reconhecimento de locutor, segue-se na Figura 1, em que um sinal de voz é passado por um filtro passa-baixa a uma frequência aproximada de 4 KHz, *anti-aliasing* satisfazendo o Teorema da Amostragem [9], posteriormente usa-se um conversor analógico digital para amostrar o sinal. São extraídas as características do sinal digital e um conjunto de matrizes é obtido após a aplicação da quantização vetorial, essas serão as características de entrada para aplicação dos classificadores polinomiais. Então determina-se os modelos de locutores,

e compara-se os padrões. A seguir o sistema toma a decisão baseando-se no "melhor score".

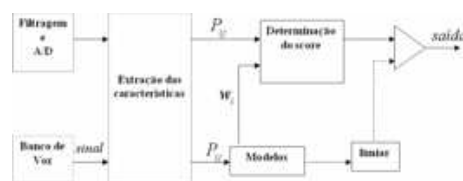


Figura 1: Modelo do reconhecimento de locutor pela voz

1.3 Característica do Sinal

Os sinais utilizados nos testes foram digitalizados a uma taxa de amostragem 11025 Hz, utilizando 16 bits e 1 canal, *Mono*, em uma placa compatível a *Sound Blaster*.

Um banco de voz foi criado para testes de reconhecimento de locutor, tipo texto dependente, usando três repetições da contagem natural de zero a nove para 30 locutores de ambos os sexos e com faixa etária entre 18 e 40 anos. O tempo médio de gravação foi de 11,43 s. e tamanho médio dos arquivos foi de 246 KBs.

2 Extração de Característica

A seqüência formada pelas amostras de voz foram divididas em quadros com 256 amostras com superposição de 50 % posteriormente aplica-se o *Janelamento de Hamming*, para que sejam eliminadas as componentes de alta frequência provenientes do janelamento [9]. O sistema calcula então 12 coeficientes de LPC - *Linear Predictive Coding* [10]. Desta forma para cada seqüência que forma o sinal original é atribuída uma matriz, A, na ordem de $w \times 12$, onde n representa o número de janelas:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,12} \\ a_{2,1} & a_{2,2} & \dots & a_{2,12} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ a_{w,1} & a_{w,2} & \dots & a_{w,12} \end{bmatrix} \quad (1)$$

Calcula-se para cada uma das matrizes obtidas os seus respectivos *codebooks* com aplicação da quantização vetorial, como mostra a Figura 2.

2.1 Quantização Vetorial

A quantização vetorial é um princípio de compressão de dados que permite baixa razão de codificação. Este é um tipo de classificador que utiliza as similaridades para classificar os padrões, isto é, pode-se aproximar um vetor x do R^2 pelo vetor

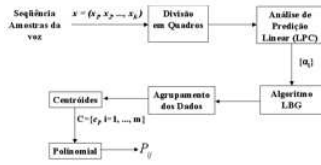


Figura 2: Processo de extração de características

mais próximo, representante de uma classe, denominado centróide [6, 11]. Este algoritmo tem como entrada de dados um conjunto de matrizes A , tal como a equação 1, o número de vetores códigos do processo, e a aproximação necessária. A saída são *codebooks*.

Mais precisamente, calcula-se o vetor médio pela equação 2:

$$c_i = \frac{1}{n} \sum_{j=1}^n a_j \quad (2)$$

onde n é o número de elementos da seqüência de A .

Assim, através de perturbações ϵ , e redivisão em classes menores da região, obtém-se um conjunto de centróides C , os *codebooks*, dado pela matriz:

$$C = \begin{bmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,12} \\ c_{2,1} & c_{2,2} & \dots & c_{2,12} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ c_{m,1} & c_{m,2} & \dots & c_{m,12} \end{bmatrix} \quad (3)$$

onde m é o número de centróides desejado.

Um quantizador vetorial simples faz a classificação dos padrões utilizando medidas de distorção entre os codebooks, em sinais de voz a mais utilizada é a proposta por Itakura [12]. Neste trabalho usa-se essa medida de distorção apenas no algoritmo LBG [13] para calcular-se a distância entre os vetores. A classificação dos padrões é feita via Classificador Polinomial.

2.2 Classificador Polinomial

Um classificador polinomial pode ser generalizado por uma função discriminante [14].

2.2.1 Classificador Hiperplano

Um hiperplano é definido como uma equação linear:

$$v_1 x_1 + v_2 x_2 + v_3 x_3 + \dots + v_n x_n = p \quad (4)$$

ou pelo produto escalar:

$$v \cdot x = p \quad (5)$$

onde o vetor $v = (v_1, v_2, \dots, v_n)^T$ é normal ao hiperplano e qualquer hiperplano H define dos semi-espacos, dados pela eq.6 e pela eq.7:

$$v \cdot x \geq p \quad (6)$$

O outro semiplano é definido pela desigualdade:

$$v \cdot x \leq p \quad (7)$$

Para passar do interior de um semiplano para o interior do outro tem-se que cruzar obrigatoriamente o hiperplano dado pela eq. 5.

2.2.2 Função Discriminante

Um outro tipo de classificador é o baseado no uso de função linear discriminante definida por

$$g(x) = w^t \cdot x + w_0 \quad (8)$$

onde w é chamado vetor de pesos e w_0 é chamado limiar de pesos.

A função discriminante $g(x)$ classifica os elementos em duas regiões ou classes W_1 ou W_2 . Se $g(x) > 0$, x pertencerá à classe W_1 e se $g(x) < 0$, x pertencerá à classe W_2 .

Para o caso de classificar-se mais de N regiões, $N \geq 2$, o processo de separação basea-se na criação de uma função discriminante para cada uma das classes, assim a eq. 8 fica generalizada como:

$$g_i(x) = w^t \cdot x + w_{i_0}, \text{ onde } i = 1 \dots N \quad (9)$$

O vetor x será classificado na classe i , se

$$g_i(x) > g_j(x), \text{ para todo } i \neq j \quad (10)$$

Um classificador polinomial pode ser generalizado contruindo-se a função discriminante. Quando a função discriminante for quadrática, tem-se:

$$f(x) = w_0 + \sum_{i=1}^N w_i x_i + \sum_{i=1}^N \sum_{j=1}^N w_{ij} x_i x_j \quad (11)$$

onde $x_i x_j = x_j x_i$.

A superfície de separação agora é uma parábola ou uma superfície hiperquadrática. A expressão pode ser generaliza, e a eq. 11 é expressa como:

$$f(x, w) = \sum_{l=0}^{N'} w_l g_l(x) \quad (12)$$

onde $g_l(x)$ é uma equação polinomial previamente definida e w_l são coeficientes a serem determinados.

A eq. 12 pode ser reescrita na forma matricial usando-se a eq.13.

$$f(x, w) = w^t p(x) \quad (13)$$

Como o vetor de entrada x são vetores formados pelas linhas da matriz C , da eq.3 faz-se a seguinte expansão polinomial para determinar-se o vetor $p(x)$, baseando se em [14].

$$p(x) = [1, x_1, x_2, \dots, x_{12}, x_1^2, x_1 x_2, x_1 x_3, \dots x_{12}^2]^t \quad (14)$$

Um outro tipo de expansão pode ser utilizado como o apresentado por [15].

Através de técnicas de análise combinatória determina-se a dimensão de $p(x)$ pela seguinte fórmula:

$$\dim(p) = \frac{n^2 + 3n + 2}{2} \quad (15)$$

onde n é a dimensão do vetor característica.

Então obtém-se uma matriz P pela aplicação da eq.14 em cada uma das linhas da matriz C de codebook.

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,m} \\ p_{2,1} & p_{2,2} & \dots & p_{2,m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ p_{k,1} & p_{k,2} & \dots & p_{k,m} \end{bmatrix} \quad (16)$$

onde m é o número de centróides utilizado e k é a dimensão do vetor $p(x)$ dado pela eq. 15.

3 Criação do Modelo

Os modelos do locutores (*speakers models*) são calculados, para cada um dos arquivos parâmetros usados no reconhecimento [16, 14], ou seja, em um projeto com N locutores, os modelos serão dados por:

$$R.w_i = M^t.O_i, \text{ onde } i=1, \dots, N \quad (17)$$

Para isto inicialmente, determina-se a matriz, M , pela seguinte equação:

$$M = \begin{bmatrix} P_{1,1} \\ P_{1,2} \\ \cdot \\ \cdot \\ P_{N,1} \\ P_{N,2} \end{bmatrix} \quad (18)$$

onde N é número de locutores parâmetros do projeto e utiliza-se dois arquivos de cada locutor para se estabelecer os parâmetros.

Determina-se então uma segunda matriz R , dado por:

$$R = M^t * M \quad (19)$$

Posteriormente encontra-se uma matriz O_i , que consiste de 1's na posição dos dados do locutor em questão e 0's nas demais. Os modelos na identificação são simplesmente calculados pela eq. 17, enquanto na verificação faz-se necessário estabelecer além dos modelos um *score* limiar para estabelecer as comparações, que é assunto da próxima seção.

4 Determinação do Score e Tomada de Decisão

Dado o sistema da fig. 3, o *score* é determinado por:

$$s_i = \frac{1}{N} \sum_{j=1}^N f(w_i, x_j) \quad (20)$$

Ou ainda, segundo a eq. 13:

$$s_i = \frac{1}{N} \sum_{j=1}^N w_i^i p(x_j) \quad (21)$$

onde: N é o número dos locutores usados no projeto, j varia de acordo com o locutor parâmetro e i varia de acordo com o locutor em teste.



Figura 3: Processo de determinação do score

Encontrados os s_i 's, o melhor *score* é escolhido. A tomada de decisão é feita de forma diferente para identificação e verificação do locutor.

No processo de identificação, um locutor teste i é identificado como sendo o locutor parâmetro i , sepre que:

$$i = j \Leftrightarrow s_{ij} \geq s_{ik}, \forall i, j, k = 1, \dots, N \text{ com } j \neq k \quad (22)$$

Na verificação é necessário determinar-se um *score* limiar [16], este trabalho propõe a seguinte equação:

$$s_{lim} = \min\{s_{spk}\} \quad (23)$$

onde, s_{spk} são os locutores aptos ao processo.

Assim pela eq. 23 um locutor i terá resposta positiva do processo, sempre que:

$$s_i \geq s_{lim} \quad (24)$$

5 Resultados

Algumas rotinas foram implementadas em *Mat-Lab* para aplicar a técnica proposta. Os resultados encontrados no processo de reconhecimento apresentam-se em duas sessões apenas por efeito didático.

5.1 Verificação

Na realização do estudo propõem-se o seguinte projeto. Em um conjunto de 30 locutores classificamos aleatoriamente um grupo de 10 locutores como aptos e um conjunto de 20 locutores como sendo inaptos a um processo de verificação. Testes envolvendo uma variação no número de centróides foram feitos e os resultados são acompanhados na tabela 1:

Tabela 1: Tabela de desempenho do sistema na Verificação.

Número de Centróides	e_A (%)	e_I (%)	e_T (%)
4	5	10	6,67
8	10	0	6,67
16	5	0	3,34

O gráfico apresentado na Figura 4 exibe o posicionamento dos scores aptos e inaptos com relação ao limiar, usando 16 centróides.

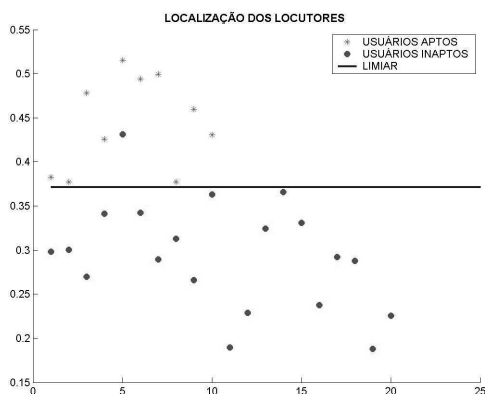


Figura 4: Processo de verificação usando 16 centróides

Nos processo de verificação podem ocorrer 2 tipos de erros, que são apresentados na Tabela 1 um na

classificação da aptidão, chamado erro de aptidão e_A , que ocorre quando o locutor está inapto mas é classificado como apto, caso ocorra o contrario denomina-se erro de inaptação e_I . E chama-se erro total, e_T o erro do processo.

5.2 Identificação

No processo de identificação apresenta-se a tabela de confiabilidade do processo na Tabela 2

Tabela 2: Tabela de desempenho do sistema na Identificação.

Número de Centróides	e (%)
4	23,34
8	13,34
16	3,34

Uma ilustração para o processo de identificação é apresentado na Figura 5 em que foi solicitado a identificação dos locutores 17 e 18 com ocorrência de acerto e erro respectivamente no processo.

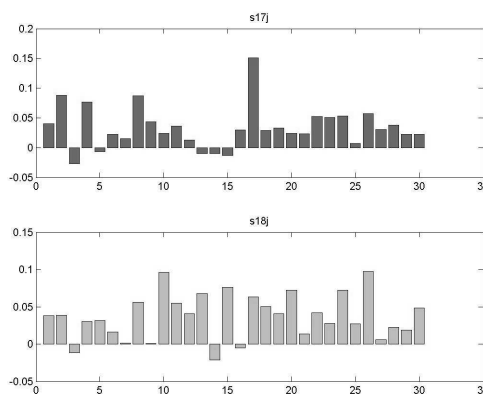


Figura 5: Processo de identificação

Outros processos de reconhecimento já propostos pela literatura pode ser visto na Tabela 3 que envolvem características LPC.

Tabela 3: Taxa erro em sistemas de reconhecimento baseado em análise de LPC

Proposto por	Método	Erro
Soong, et al. [6]	QV	5%
Figueiredo [11]	QV	15%
Tishby 1991 [4]	HMM(AR Mix)	2.8 %

As taxas de erro obtidas pelo método usado por *Soong* em [6] e *Figueiredo* em [11] tratam-se apenas para identificação. O erro do método usado por *Tishby* em [4] refere-se a verificação.

6 Conclusões

Os classificadores polinomiais são técnicas baseadas nos métodos determinísticos, portanto um alinhamento temporal é normalmente necessário. Porém com a combinação desses classificadores com a quantização vetorial, usando os codebooks gerados, o processo de alinhamento passa ser dispensável.

É possível notar também que este sistema envolvendo combinações de duas técnicas, quantização vetorial e classificação polinomial (Tabelas 1, 2) obteve maior rendimento a alguns dos métodos apresentados na Tabela 3, em ambos processos de reconhecimento. Neste trabalho apresenta-se projetos com codebooks usando no máximo 16 centroídes, que é consideravelmente pequeno quando se utiliza a quantização vetorial simples (sem composição com outra técnica) para um processo de reconhecimento de locutor.

A exploração dos classificadores polinomiais aplicado ao reconhecimento de locutores pela voz, é uma técnica recente. Porém ainda há muito a ser explorado, principalmente no que tange a combinação com outras técnicas.

Abstract. The Speaker Recognition by voice is a high complexity problem. This paper proposes a method to solve this problem by using polynomial classifiers and vectorial quantization. With the application of the vector quantization, the samples of the voice signal do not need to be time aligned. The vector quantization produces a set of matrices that will be the input of the system based on polynomial classifiers. The polynomial classifiers have excellent properties as classifiers. Because of the Weierstrass theorem, polynomial classifiers are universal approximators to the optimal Bayes classifier. It will be shown that with the combination of those two methods the system reached a better performance than with the single vector quantization for the speaker recognition.

Referências

- [1] A.E. Rosemberg, *Automatic Speaker Verification: A Review*. Proceedings of the IEEE, p.475-487, v.64, n.4, Abril 1987.
- [2] B.S. Atal, *Automatic Recognition of Speakers from Their Voices*. Proceedings of the IEEE, p.460-474, v.64, Abril 1976.
- [3] R.D. Peacocke e D.H. Graf, *An Introduction to Speech and Speaker Recognition*. Computer, p.26-33, Agosto 1990.
- [4] N.Z. Tishby, *On Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition*. IEEE Trans. Acoust., Speech, Signal Processing, p.563-570, v.39, n.3, 1991.
- [5] J.P. Campbell Jr., *Speaker Recognition: A Tutorial*. Proceedings of the IEEE, p.205-212, v.85, n.9, Setembro 1997.
- [6] F.K. Song, A.E. Rosemberg, B.H. Juang e L.R. Rabiner, *A Vector Quantization Approach to Speaker Recognition*, in Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Tampa, FL, 1985, p. 387-390.
- [7] K.R. Farrell, R.J. Mammone e K.T. Assaleh, *Speaker Recognition using Neural Networks and Convencional Classifiers*. IEEE Trns.Speech Audio Processing, p.194-205, v.2, Janeiro 1994.
- [8] K.T. Assaleh e W.M.Campbell, *Speaker Identification Using a Polynomial-Based Classifier*. Fifth International Symposium on Signal Processing and Its Applications, Brisbane, Austrália, p.115-118, v.22-25, Agosto 1999.
- [9] B.P. Lathi, *Linear Systems and Signals*, New York: Oxford University Press, Inc., 2005. p.975.
- [10] L.R. Rabiner e R.W. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prince Hall, Inc., 1978. p.512.
- [11] S.A. Figueiredo, *Contribuição ao Estudo de Reconhecimento do Locutor pela Voz*. Uberlândia: Universidade Federal de Uberlândia, Faculdade de Engenharia Elétrica, 1990.
- [12] J. Makhoul, S. Roucos e H. Gish, *Vector Quantization in Speech Coding*, *Proceedings of the IEEE*, p.1551-1558, v.73, n.11, Novembro 1980.
- [13] Y.Linde, A.Buzo e R.M. Gray, *An Algorithm for Vector Quantizer Design*. IEEE Transaction on Communications, p.84-85, v. COM28, n.1, Janeiro 1980.
- [14] W.M.Campbell e K.T. Assaleh, *Speaker Recognition with Polynomial Classifiers*. Proceedings of the IEEE, p.205-212, v.10, Maio 2002.
- [15] V. Wan e S. Renals, *Evaluation of Kernel Methods for Speaker Verification and Identification*. Proceedings of the IEEE, p.669-672, 2002.
- [16] W.M.Campbell e K.T. Assaleh, *Polynomial Classifier Techniques for Speker Verification*. Proceedings of the IEEE, p.321-324, Maio 1999.