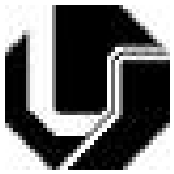
UNIVERSIDADE FEDERAL DE UBERLÂNDIA

FACULDADE DE ENGENHARIA ELÉTRICA

PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA



RECONHECIMENTO DE LOCUTOR PELA VOZ USANDO O CLASSIFICADOR POLINOMIAL E QUANTIZAÇÃO VETORIAL

WEMERSON DELCIO PARREIRA

ABRIL

2005

ERRATA(S)

Após impressão do presente trabalho, detectou-se na leitura de revisão as folhas editoriais citadas a seguir:

1. Capítulo, Página, Parágrafo/ Tabela/ Figura/ Equação:
    - Onde se lê: “”
    - Leia-se: “”

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

FACULDADE DE ENGENHARIA ELÉTRICA

PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

RECONHECIMENTO DE LOCUTOR PELA VOZ USANDO O CLASSIFICADOR POLINOMIAL E QUANTIZAÇÃO VETORIAL

Dissertação apresentada à Universidade Federal de Uberlândia por Wemerson Delcio Parreira, como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica, aprovada em 01/ 04/ 2005 pela seguinte Banca Examinadora:

Prof. Dr. Antônio Cláudio P. Veiga, UFU.
Prof[a] Dr[a] Edna Lúcia Flores, UFU.
Prof. Dr. Gilberto Arantes Carrijo, UFU (Orientador).
Prof. Dr. João Cândido Lima Dovicchi, UFSC.

RECONHECIMENTO DE LOCUTOR PELA VOZ USANDO O CLASSIFICADOR POLINOMIAL E QUANTIZAÇÃO VETORIAL

WEMERSON DELCIO PARREIRA

Dissertação apresentada por Wemerson Delcio Parreira à Universidade Federal de Uberlândia, como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica.

Prof. Gilberto Arantes Carrijo, Dr. Orientador	Prof. Gilberto Arantes Carrijo, Dr. Coordenador do Curso de Pós Graduação

Aos meus amados pais,

Antônio e Rosa,

e em memória de meus queridos avós,

Odilon , Maria de Jesus, Frederico ,

Leonel e Maria Delfina .

AGRADECIMENTOS

A Deus, pelas bênçãos diárias.

Ao Prof. Dr. Gilberto Carrijo, pela orientação, confiança e apoio.

Ao Prof. Dr. João Dovicchi pela confiança em minha competência, respeito, amizade, apoio e incentivo nos momentos críticos.

Ao meu querido pai, Antônio pelo amor e à minha querida mãe Rosa pelo carinho, amor e companheirismo em todos os momentos, aos meus adoráveis irmãos Anderson e Wesley pelo exemplo, apoio e compreensão e a minha querida madrinha Aparecida pela presença e dedicação.

Aos companheiros de curso, em especial, ao Anderson que muito contribuiu com proveitosas dicas e longas discussões.

Aos professores e funcionários da COPEL, em especial, Prof ª. Drª Edna Lúcia, pelo apoio e atenção e Dna. Marli, pela colaboração em questões administrativas e acadêmicas.

Ao CNPq, pelo apoio financeiro durante o período em que transcorreu este trabalho.

Aos meus professores da Faculdade de Matemática que me promoveram o início do meu desenvolvimento crítico científico.

Àqueles que gentilmente cederam sua voz para a criação do Banco de Dados: Anderson, Milena, Rodrigo, Karine, Arquimedes, Alexandre, Asmara, Púbio, Gláucia, Marlene, Edgar, Adeon, Leonardo, Marcos, Renata, Andreza, Daniela, Marta, Raquel, Marlene Roque, Saulo, Éder, Heron, Maria Terezinha, Ana Marcela, Ronaldo, Lidiane, Moline e Cristiane.

À minha querida e doce Marli S. Santos pelo apoio na fase conclusiva deste trabalho.

E a todos aqueles que, direta ou indiretamente colaboram para que este trabalho pudesse ser realizado.

RESUMO

Parreira, W. D. Reconhecimento de Locutor pela Voz Usando o Classificador Polinomial e Quantização Vetorial, Uberlândia, UFU, 2005, 142 p.

O Reconhecimento de locutores pela voz é um problema de alta complexidade. Um método que utiliza classificadores polinomiais e quantização vetorial será proposto neste trabalho para a solução deste problema. Inicialmente é necessário conhecer o processo de produção da voz humana e as características necessárias para o reconhecimento assim como as condições que afetam tais características. Foi criado um banco de dados contendo arquivos de voz digital, que foram divididos em quadros com superposição e aplicação de janelamento do *Hamming*. Define-se um sistema físico-matemático e aplica-se a Quantização Vetorial aos arquivos do banco. Posteriormente são calculados os coeficientes LPC de cada quadro para se criar os *codebooks* que serão os vetores característica. Um algoritmo é treinado para se criar a estrutura classificatória usando expansão polinomial. São gerados, então, os modelos dos locutores. Um *score* é atribuído a cada um dos locutores e são feitas as devidas comparações obtendo-se assim o reconhecimento. Finalmente, estabelece-se uma comparação entre alguns métodos de reconhecimento de locutores já propostos.

Palavras - Chaves: Quantização Vetorial, Classificadores Polinomiais, Reconhecimento do Locutor, Identificação do Locutor, Verificação do Locutor.

ABSTRACT

Parreira, W. D. Voice Speaker Recognition using a Polynomial Classifier and Vector Quantization, Uberlândia, UFU, 2005, 142 p.

The speaker recognition by voice is a high complexity problem. To solve this problem a method using polynomial classifiers and vector quantization will be presented in this work. Initially it is necessary to know the production process of the human voice and the necessary characteristics to the recognition as well as the conditions that affect these characteristics. A voice data base containing digitally sampled files have been created. These files where splited in frames with overposition and to them were applied the *Hamming* window. A physical-mathematical system is defined and vector quantization is applied to the samples from the database. The LPC coefficients are calculated to each frame to create the codebooks that will be the characteristic vectors. An algorithm is trained to create the classifier structure using polynomial expansion. Then, the speaker models are created. A score is assigned to each speaker and the proper comparisons are done resulting in the recognition. Finally is established a comparison between some recognition methods already presented.

Keywords: Vector Quantization, Polynomial Classifiers, Speaker Recognition, Speaker Identification, Speaker Verification.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS E UNIDADES

ACRÔNIMOS

A/D	- Analógico/ Digital
ASI	- Automatic Speaker Identification (Identificação Automática de Locutor).
ASR	- Automatic Speaker Recognition (Reconhecimento Automático de Locutor).
AVS	- Automatic Verification Speaker (Verificação Automática de Locutor).
DTW	- Dynamic Time Warping
HMM	- Hide Markov Model
LPC	- Linear Predictive Coding
mse	- Minimum squared error
QV	- Quantização Vetorial

UNIDADES

Hz	Hertz (ciclos por segundo)
KB	Kilo bytes
Kbpz	Kilobits por segundo
s	segundos

CAPÍTULO I

O RECONHECIMENTO E VERIFICAÇÃO DO LOCUTOR PELA SUA VOZ

1.1 - 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 sua voz é uma técnica que teve início a mais de 30 anos [2], [25]. Mais 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 [21] onde a porcentagem de reconhecimento correto pode chegar até 99%.

Em várias aplicações de reconhecimento da fala é muito difícil ter um desempenho que se aproxima do ser humano, mas no reconhecimento automático do locutor (ASR) o sucesso das máquinas é superior ao do homem.

Hoje, no entanto várias pesquisas estão direcionadas para entender como uma pessoa distingue um locutor dentre vários outros. Porque o reconhecimento do locutor é possível? Quais são as variações entre os locutores e como elas se manifestam em termos de níveis acústicos?

Estas respostas, entretanto não são fáceis de responder. A maneira como se reconhece um locutor pode ser apresentado de duas maneiras:

- Verificação automática do locutor (ASV) e
- Identificação automática do locutor (ASI).

A verificação automática do locutor (ASV) é usada para verificar se a pessoa que reivindica a verificação é realmente a pessoa, e não um impostor. Isto pode acontecer quando uma pessoa digita um código e logo em seguida fala uma frase. A verificação tem por finalidade verificar se a voz é realmente da pessoa que é proprietária do código ou se é um impostor.

A ASV é uma tarefa mais simples, pois ela compara um padrão teste com um padrão de referência que envolve uma decisão binária, ou seja, a resposta será ‘sim’ ou ‘não’.

O processo de identificação automática do locutor (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. A ASI 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, 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 (intralocutor) 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 identificação é realizada com a comparação de um texto falado no momento do reconhecimento com outro previamente gravado pelos locutores.

Este capítulo apresenta as diferenças entre os sistemas de Verificação (ASV) e Sistemas de Identificação (ASI). São apresentados também os Métodos de Verificação e Identificação que são os classificados em Estatísticos e Determinísticos. São descritos os bancos de dados mais usados atualmente nas pesquisas envolvendo sinais de fala. E finalmente são feitas considerações finais sobre este capítulo.

1.2 - Formulação do Problema e Aquisição de Dados

Um sistema ASR consiste de cinco fases: filtragem e digitalização do sinal, extração das características, comparação com padrões, tomada de decisão, e o registro para gerar o modelo de referências, como mostrado na Figura 1.1.

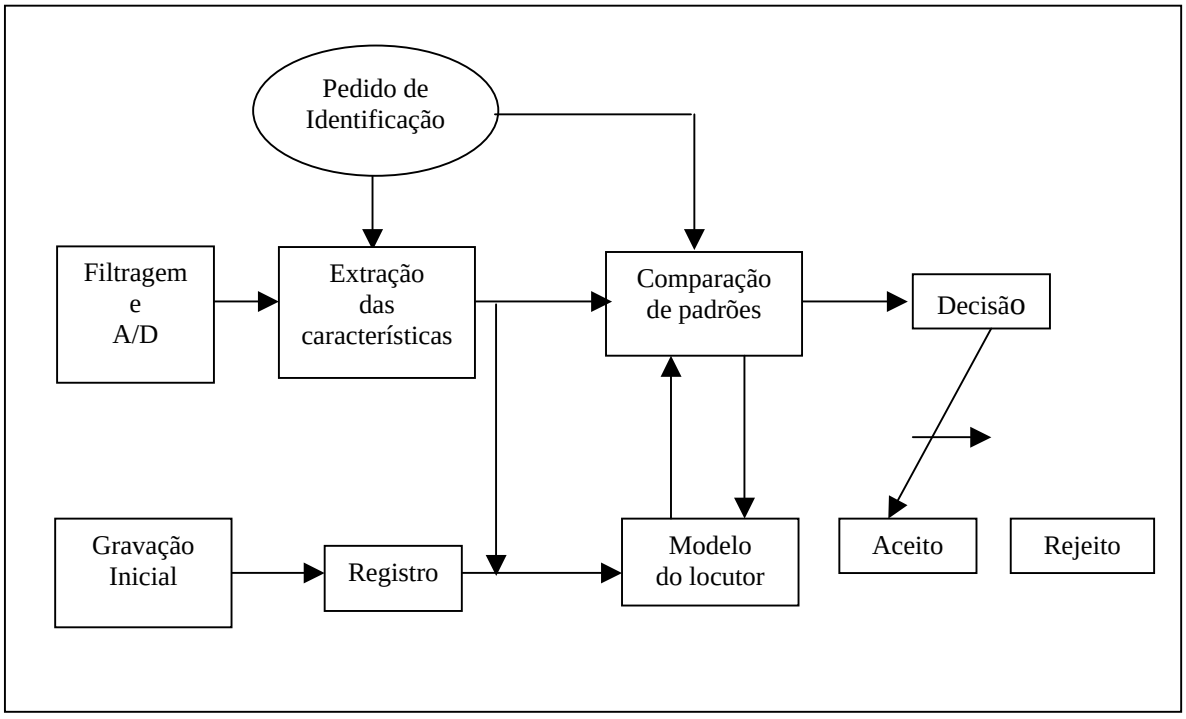


Figura 1.1- Sistema de verificação

a- Na filtragem e digitalização o sinal é passado por um filtro passa baixa na freqüência de corte normalmente de 4KHz, também chamado filtro *"anti-aliasing"*. Em seguida o sinal é amostrado e digitalizado por um conversor analógico para digital (A/D). A freqüência de amostragem precisa ser superior ou igual a duas vezes a freqüência de corte do filtro, segundo o teorema da amostragem.

b- No processo de extração das características, o sinal é dividido em quadros com duração que pode variar aproximadamente de 10 ms até 30 ms. Em cada quadro extrai os parâmetros que vão caracterizar o quadro. Os parâmetros mais comuns são: os coeficientes de Fourier, os coeficientes LPC, a saída de um banco de filtros e etc. Para uma freqüência de amostragem de 8 KHz, tem-se um intervalo entre as amostras de 0,125 ms. Um total de 256 amostras fornece um intervalo do quadro de 32ms.

c- No processo de comparação, os vetores das características do locutor em teste são comparados com os padrões armazenados de todos os locutores, para tomar a decisão de identificação ou verificação.

d- Quando se trabalha com ASV, a tomada de decisão é feita baseando se em um limiar previamente estabelecido, sendo aceito ou não o locutor que pede para ser verificado. No sistema ASI é escolhido um locutor dentro da população de todos os locutores tal que os vetores das características em teste mais se aproximam.

e- A geração de um modelo de referência é feita no estágio inicial, a partir das características de cada locutor em treinamento. Para isso usa se frases previamente escolhidas. O tempo de gravação de cada locutor para a extração das características varia de acordo com a quantidade de locutores a serem identificados ou verificados no processo.

1.3 - 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 é necessário para calcular a distância.

A distância entre dois quadros é calculada pela distância de Mahalanobis [4],

$$d(x_i,x) = (x_i - x)^T W(x_i - x) \qquad (1.1)$$

onde W é a matriz de pesos, x é o vetor que representa um locutor e x_i é o vetor de entrada que se deseja classificar. Para o caso da distância Euclidiana W é uma matriz identidade.

Dentre os principais métodos determinísticos tem-se:

a- Dynamic Time Warping (DTW)

O método mais popular para compensar a variação da velocidade da voz de uma pessoa quando se pronuncia uma mesma frase é a técnica chamada DWT[4]. Uma seqüência de vetores amostras de um padrão teste (x_1, ..., x_N) precisa ser comparada a uma seqüência de amostras de um padrão referência (y_1, ..., y_M), ambas provenientes do mesmo texto, mas faladas com velocidades diferentes. A Figura 1.2 mostra um exemplo de alinhamento no tempo de um padrão teste com um padrão [28].

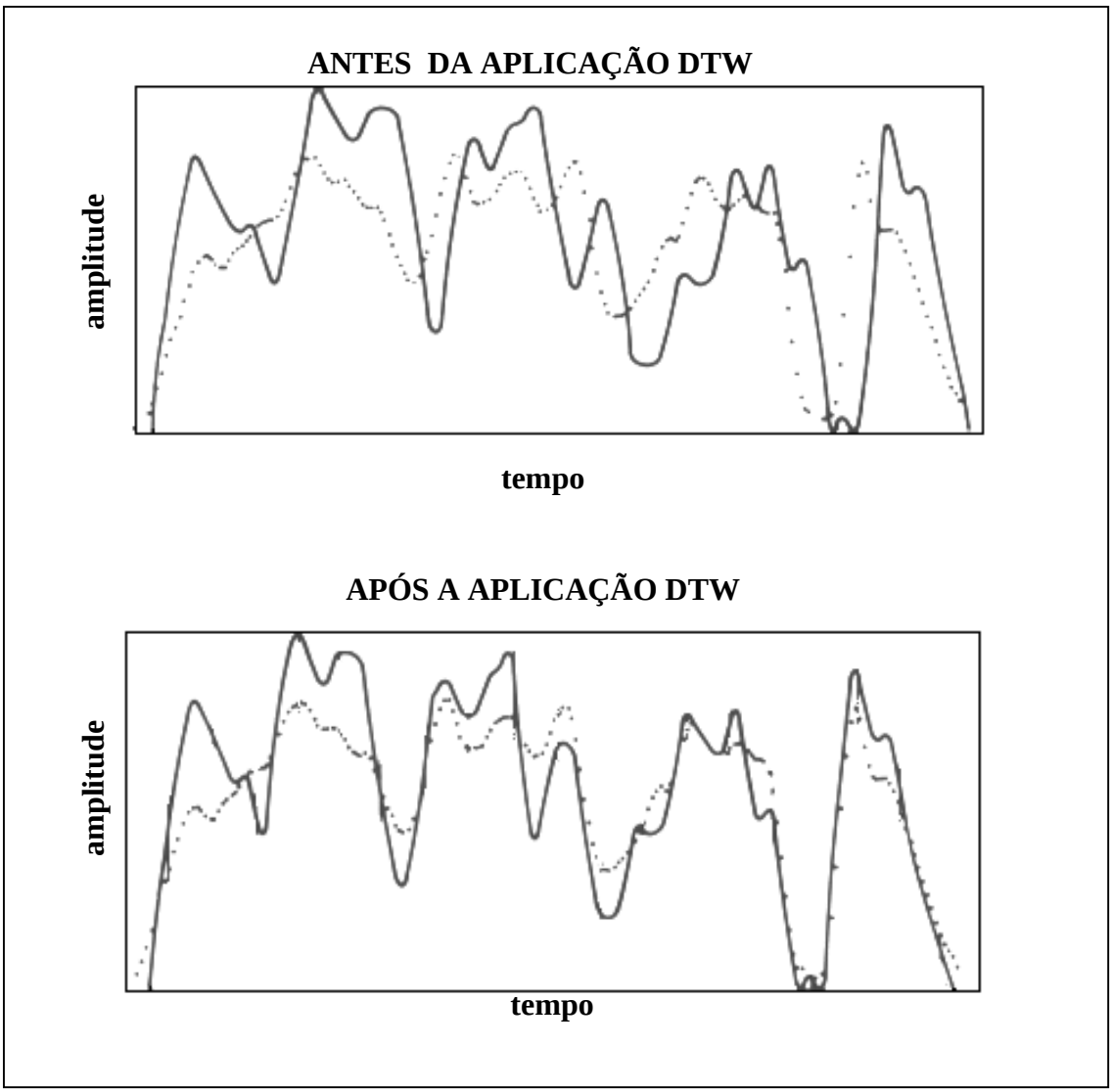


Figura 1.2 – Aplicação da DTW. [25]

b- Quantização Vetorial

A quantização vetorial usa um conjunto de vetores para construir um *"codebook"*, com um número *N* de centróides. O processo de treinamento usualmente usa o algoritmo conhecido de LBG (Lindle, Buzo e Gray) [17] para treinamento e construção do *codebook*. O número de centróides pode variar de acordo com a aplicação. Cada vetor característico do padrão teste de um quadro é comparado aos centróides do *codebook*. A aquele que produzir a menor distância é escolhido para representar o vetor característico.

c- Redes Neurais

Esta técnica exige um treinamento inicial do conjunto de dados a serem treinados pela rede, mas para classificar é necessário usar um alinhamento tal como o DTW.

d- Técnica Polinomial

A técnica polinomial [1], que é o objetivo principal deste trabalho consiste em calcular um valor de saída (média) para os vetores característicos em teste proveniente de um locutor que se deseja reconhecer, como mostrado na Figura 1.3.

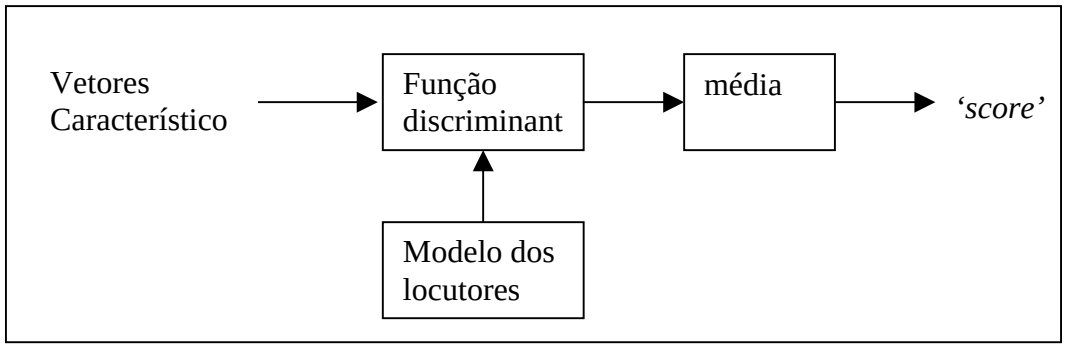


Figura 1.3 - Estrutura do classificador polinomial

Os vetores característicos são obtidos da fala do locutor que se deseja reconhecer. Os modelos dos locutores estão armazenados e o objetivo é achar qual modelo se aproxima mais dos vetores característicos de entrada. Estes vetores x_1, ..., x_N são processados por um discriminador polinomial da forma,

$$f(x, w) = w^t p(x) \tag{1.2}$$

onde, $p(x)$ é um vetor polinomial com funções base da ordem K, w^t é o vetor peso de um determinado locutor, e x é um vetor característico de um quadro. Calcula-se a média para

todos os vetores característicos de um mesmo locutor. Repete-se o processo para todos os locutores. O locutor que produzir a maior média será escolhido como sendo o representante dos vetores característicos de entrada.

Os métodos estatísticos são:

a- Função Densidade de Probabilidade

Para cada locutor é determinada uma função densidade de probabilidade que o representa. O aceite ou a rejeição de um locutor no processo de verificação é determinado pela função verossimilhança.

b- Hide Markov Model (HMM)

No HMM os sinais são considerados como um modelo de Markov, e cada estado corresponde a um evento. Constrói-se uma HMM para a cadeia apartir de observações onde se calcula a probabilidade de transição entre os estados $a_{ij} = p(s_i/s_j)$.

Em alguns casos usa se a combinação de duas ou mais das técnicas usadas acima como é o alvo deste trabalho.

1.4 - *Corpus* para Reconhecimento do Locutor

O uso de *corpus* de sinais da fala para o desenvolvimento do reconhecimento de locutor é um fator importante e isto tem sido feito nos últimos 10 anos. O principal benefício ao usar um padrão é a facilidade de comparações entre as várias técnicas desenvolvidas pelos pesquisadores. Os principais *corpus* de sinais de voz são:

a- TIMIT

O TIMIT foi desenvolvido para proporcionar um conjunto de dados para estudos de sistemas de reconhecimento da fala e sistemas de reconhecimento de locutores. Ele inclui: o FFTMTIMIT que foi gravado via microfone, o NTIMIT que foi gravado usando uma linha telefônica e o CTIMIT que foi gravado usando um celular movimentado dentro de uma Van [5].

A tabela 1.1 mostra a descrição do *corpus* TIMIT.

TABELA 1.1 – *Corpus* TIMIT.

Número de locutores	630
Número de sessões	1
Intervalo entre as sessões	nenhum
Tipo de fala	leitura de sentenças
Microfone	fixo e de faixa larga

b- O SIVA

O SIVA é um corpus italiano para a identificação e verificação do locutor, composto de uma coleção de 2000 chamadas de um telefone público. Ele é constituído de voz masculina e feminina de usuários e impostores [5].

A tabela 1.2 mostra a descrição do *corpus* SIVA

TABELA 1.2 – *Corpus* SIVA.

Número de locutores	671

Número de sessões/locutor	1-26
Intervalo entre as sessões	dias-meses
Tipo de fala	palavras, dígitos e perguntas curtas
Microfone	telefones variáveis

c- PolyVar

O PolyVar compreende um corpus nativo e não nativo de locutor em francês. Ele consiste da leitura da fala espontânea de 160 horas de voz [5].

A tabela 1.3 mostra a descrição do *corpus* PolyVar.

TABELA 1.3 – *Corpus* PolyVar.

Número de locutores	143
Número sessões/locutor	1-229
Intervalo das sessões	dias-meses
Tipo de fala	dígitos, sentenças, questões e fala espontânea
Microfone	telefones variáveis

d- O POYCOST

O POYCOST foi coletado sob um projeto europeu o COST-50 para a verificação de locutores [5]. A maioria dos locutores foram de não nativos, cobrindo 13 países da Europa. A fala foi coletada de uma linha telefônica.

A tabela 1.4 mostra a descrição do *corpus* POYCOST.

TABELA 1.4 – *Corpus* POYCOST.

Número de locutores	133

Número sessões/locutor	>5
Intervalo das sessões	dias-semanas
Tipo de fala	dígitos, sentenças e monólogo livre
Microfone	telefones variáveis

e- O KING

Este corpus foi coletado pela ITT nos USA em 1987[5].

A Tabela 1.5 mostra a descrição do *corpus* KING.

TABELA 1.5 – *Corpus* KING.

Número de locutores	51 (homens)
Número sessões/locutor	10
Intervalo das sessões	dias-meses
Tipo de fala	espontânea descrição de locutor
Microfone	telefones variáveis

f- YOHO

O YOHO foi desenvolvido na Alemanha com sistema dependente do texto para verificação do locutor, gravado em sistema de baixo ruído[5].

A Tabela 1.6 mostra a descrição do *corpus* YOHO.

TABELA 1.6 – *Corpus* YOHO.

Número de locutores	138

Número sessões/locutor	4 registros, 10 verificação
Intervalo das sessões	dias-meses
Tipo de fala	frases de dígitos
Microfone	fixo e de alta qualidade

1.5 - Sistemas de Reconhecimento e Estado de Arte

Existe um número considerável de sistemas de reconhecimento do locutor pela fala nas indústrias e laboratórios. Os principais centros que tem pesquisado na área são: AT&T [4], ITT [4], MIT [4], National Tsing Ha University (Taiwan) [4], Nagoya University (Japan) [4], etc.. A Tabela 1.7 mostra um resumo de como evolui o processo de reconhecimento do locutor.

Tabela 1.7 – Evolução do processo de reconhecimento de locutor.

Fonte	Órgão	Característica	Método	Entrada	Texto	População	Erro
Atual 1974	AT&T	Cepstrum	Comparação de Padrões	Lab	Dependente	10	2%
Markel e Davis 1979	STI	LP	Estatística	Lab	Independente	17	2%
Furui 1981	AT&T	LP	Comparação de padrões	Telef	Independente	10	0,2%
Che e LI 1995	Rutgers	Cepstrum	HMM	Telef	Dependente	138	0,56%
Reynolds 1996	MIT	Mel-Cepstrum	HMM	Telef	Independente	416	11%

1.6 – Considerações Finais deste Capítulo

Neste capítulo foram apresentadas as diferenças entre os Sistemas de Verificação (ASV) e os Sistemas de Identificação (ASI). Estas diferenças motivaram o desenvolvimento de dois sistemas distintos para obtenção dos resultados deste trabalho, que será visto nos próximos capítulos.

Foram apresentados também os métodos de Verificação e Identificação, que são classificados em estatísticos e determinísticos. Este trabalho faz uso da combinação de duas técnicas determinísticas.

Finalmente a descrição do banco de dados mais usados atualmente nas pesquisas envolvendo sinais de fala.

CAPÍTULO II

MODELO DE REPRODUÇÃO DA VOZ

2.1 - Introdução

A fala humana torna possível a expressão e a comunicação de pensamentos, necessidades e emoções da vocalização em forma de palavras. Em adição a capacidade para a produção de sons pela laringe, que não apenas o homem possui, a fala requer um sistema de ressonância para modular e amplificar o som e um processo de articulação para transformação dos sons em palavras.

O processo de produção da fala envolve os centros da fala no cérebro, o centro respiratório na base do cérebro, o sistema respiratório, a cavidade torácica, as estruturas da laringe, a faringe, as cavidades nasais, o nariz, as estruturas e partes da boca e músculos faciais [22].

Neste capítulo é aborda a fisiologia dos órgãos responsáveis pelo mecanismo da fala. Apresenta-se algumas características importantes do aparelho vocal humano [25]. Bem como são mostrados os modelos físicos da produção da fala. E no final deste capítulo

é descrito um modelo digital de sinais de fala de acordo com [8] e são realizadas as considerações finais deste capítulo.

2.2 - Anatomia e Fisiologia da Produção dos Sons [22]

Uma visão simplificada do mecanismo de produção da fala pode ser verificada na Figura 2.1, em que os órgãos da fala estão divididos em três grupos principais: os pulmões, a laringe, e trato vocal. Os movimentos dos pulmões são responsáveis pela geração do sinal, provendo uma corrente de ar que atravessa a laringe para a produção da fala. A laringe modula o ar vindo dos pulmões e providencia uma divisão dessa corrente de ar que alimenta o terceiro órgão do grupo, o trato vocal. Este consiste de cavidade oral, nasal e faringiana, modulando a corrente de ar dando a ela uma “forma”. O som pode ser gerado por constrição ou limitação, que é feito dentro do próprio trato vocal, resultando na adição do ruído e da fonte periódica, uma fonte impulsiva.

As fontes sonoras são consideradas, impulsivas ou ruído (branco) e podem acontecer na laringe ou no trato vocal [22].

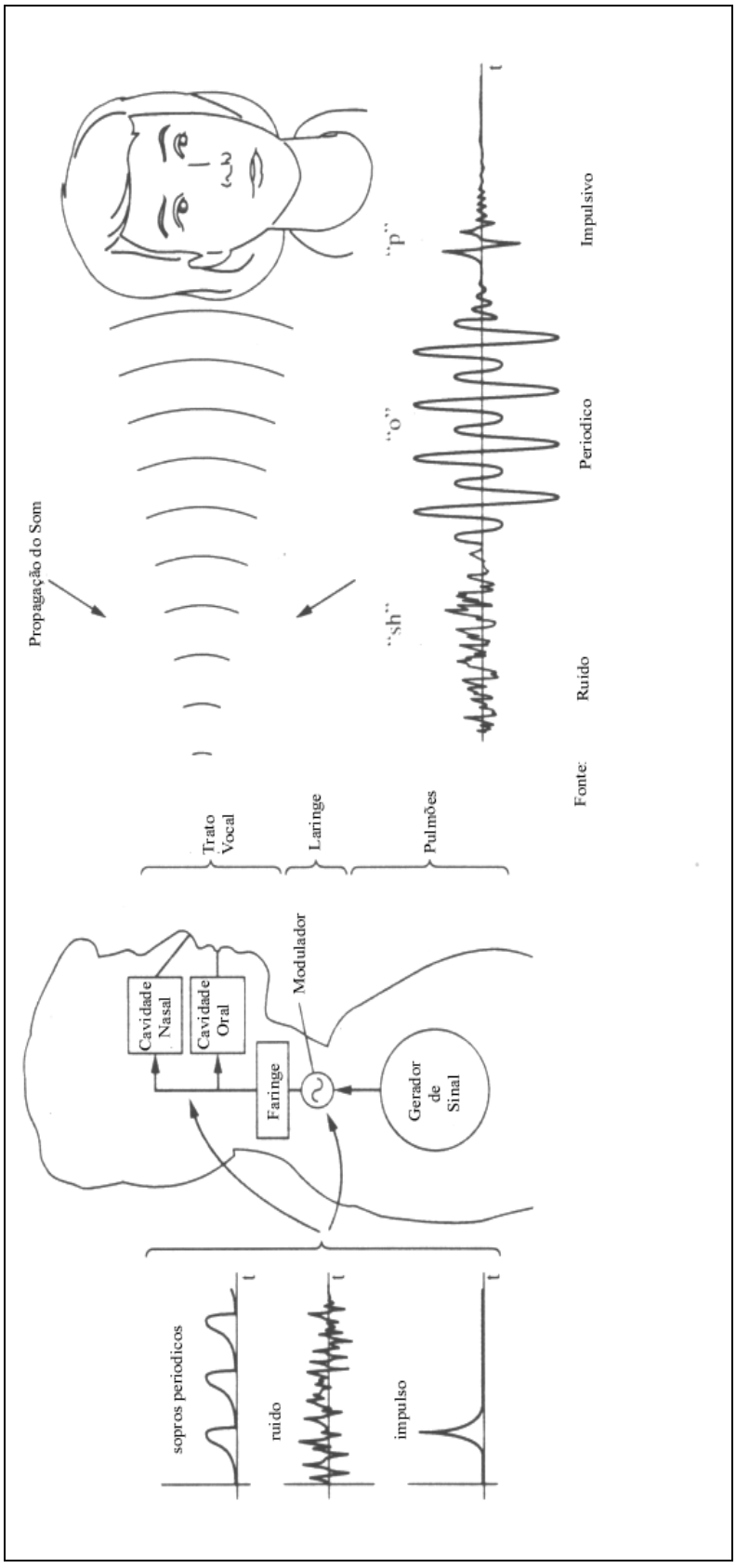


Figura 2.1 – Visão simples da produção da fala. [22]

A Figura 2.2 mostra uma visão mais real da anatomia da produção da fala a apresentada na Figura 2.1. Na Figura 2.2 é possível ver detalhes dessa anatomia, bem como associar a fisiologia e a sua importância na produção da fala.

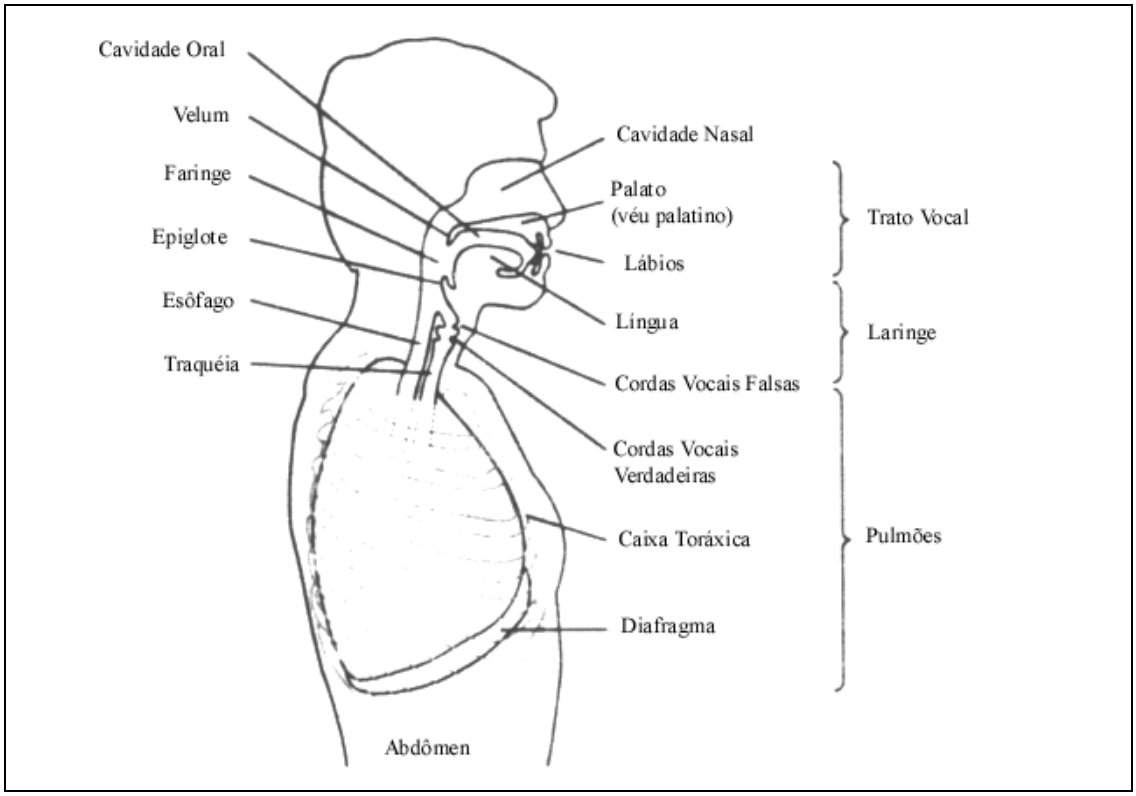


Figura 2.2 – Aparelho vocal humano [22].

2.2.1- Os pulmões

Um das funções dos pulmões é a inspiração e expiração de ar. Quando o ar é inalado, alarga-se a cavidade torácica, pelo distanciamento das costelas que protegem os pulmões e pela movimentação do diafragma que se situa abaixo dos pulmões e separa os

pulmões do abdômen; esta ação reduz a pressão de ar nos pulmões, causando uma pressão sobre a base do trato vocal e nos pulmões logo abaixo da traquéia.

A traquéia, é um órgão que possui 12 cm de comprimento e diâmetro de 1,5 a 2 cm ligando os pulmões a epiglote. A epiglote é um pequeno órgão que funciona como “interruptor” quando o ser humano se alimenta, enviando o que foi ingerido direto ao esôfago. Quando a pessoa expira, reduz o volume da cavidade torácica contraindo a musculatura dessa caixa, aumentando assim a pressurização de ar nos pulmões. Esse aumento provoca um fluxo de ar da traquéia para a laringe. Na respiração ritmicamente inspira-se para “absorver” o oxigênio e expira-se para liberar o dióxido de carbono.

Durante a fala, de outra forma, usa-se pequenos “jatos de ar” e libera continuamente controlando com a musculatura da caixa torácica. E quebra o ritmo da liberação para finalizar uma sentença ou uma frase. Durante esse tempo de expiração, a pressão do ar nos pulmões é mantida aproximadamente em nível constante, ligeiramente acima da pressão atmosférica, pelo enrijecimento ou pela contração lenta da musculatura do abdômen, apesar de variar em torno dessa devido às propriedades da variação temporal da laringe e do trato vocal.

2.2.2- A laringe

A laringe conceitualmente é um arquebouço esqueleto membranoso com as funções de respiração, fonação e proteção das vias aéreas. Sua origem é mista, endodérmica e mesodérmica. No adulto a laringe tem cerca de 5 cm de comprimento, isso para o sexo masculino sendo um pouco menor no sexo feminino. Considerado um sistema complexo de cartilagens, músculos e ligamentos; ela possui como função principal no mecanismo da fala, o controle das cordas vocais ou pregas vocais [12].

As pregas vocais são duas estruturas, formadas por ligamentos e músculos, que alongam-se desde a parte anterior até a parte posterior, como ilustra a Figura 2.3. Elas possuem 15mm de comprimento nos homens e 13 mm de comprimento nas mulheres. O orifício na forma de uma rachadura entre as pregas é denominado glote, as pregas vocais são fixas na parte frontal da laringe onde elas estão fixas a cartilagem tireóide estacionária. Esta cartilagem, a maior das cartilagens laringianas, está localizada da parte frontal (ou "pomo de Adão") às laterais da laringe. As pregas vocais são livres para movimentarem-se na parte posterior e lateral da laringe. Elas são fixadas por duas cartilagens aritenóideas que movimentam de forma deslizante na parte posterior da laringe ao longo da cartilagem cricóidea.. A dimensão da glote é controlada em parte pela cartilagem aritenóidea, e em parte pela musculatura interna das pregas vocais. Outra propriedade importante das pregas vocais, em adição a da glote é sua tensão. A tensão é controlada primariamente por um músculo das pregas vocais como também pela cartilagem que as envolve. Essas pregas, como também a epiglote, fecham durante a alimentação, isso provem um segundo mecanismo de proteção. As pregas vocais falsas, acima pregas vocais verdadeiras Figura 2.2 , fazem a terceira proteção. Elas também estendem desde o "Pomo de Adão" até as aritenóides, podem estar fechadas e vibrar, mas provavelmente abriram durante a produção da fala [10]. Foi visto que existe uma tripla barreira provida a traquéia pela ação da epiglote, dobra vocal falsa e dobra vocal verdadeira, todos as três ficam fechados durante a alimentação e abertas durante a respiração.

As aritenóides têm a forma de uma pirâmide de base triangular. A partir da base tem-se os processos musculares e vocais. Elas articulam-se como borda superior da cartilagem cricóide, em uma junta sinovial. Essa articulação é a que participa mais ativamente das funções de respiração e fonação, devido às mudanças na posição das pregas vocais. A principal propriedade dessa cartilagem vem de sua massa reduzida, com pequeno movimento de inércia ela permite a articulação das palavras. As pregas epiglóticas

inserem-se no ápice das aritenóides, como ligamento vestibular, enquanto as pregas vocais inserem-se na porção medial da base (processo vocal). O processo muscular contém inserção dos cricoaritenóides lateral e posterior.

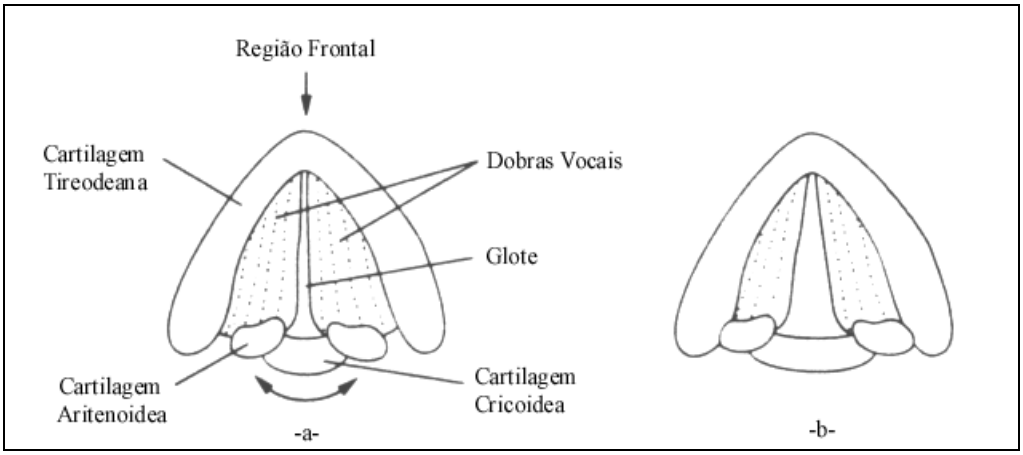


Figura 2.3 – Esboço com corte da visão superior da laringe humana: (a) expressão da voz e (b) respiração [28].

Existem três estados primários das dobras vocais: respiração, "*voiced*", "*unvoiced*". No estado de respiração na Figura 2.3b, a cartilagem aritenóidea é sustentada externamente mantendo uma abertura glotal ampla, e os músculos internos das pregas vocais relaxados. No estágio "*unvoiced*", ar proveniente dos pulmões flui livremente pela glote com obstáculo desprezível através das pregas vocais. Por outro lado, durante a fala uma obstrução da corrente de ar é provida pelas pregas. Por exemplo, durante o estado "voiced", na expressão de uma vogal, a cartilagem aritenóidea orienta-se uma a outra na Figura 2.3a. As dobras vocais tencionam e se fecham. Este fechamento parcial da glote é acentuado pela tensão das dobras causada pela sustentação própria e pela oscilação das pregas. Pode-se descrever como essa oscilação ocorre em três passos [12], esquematizado na Figura 2.4 (a).

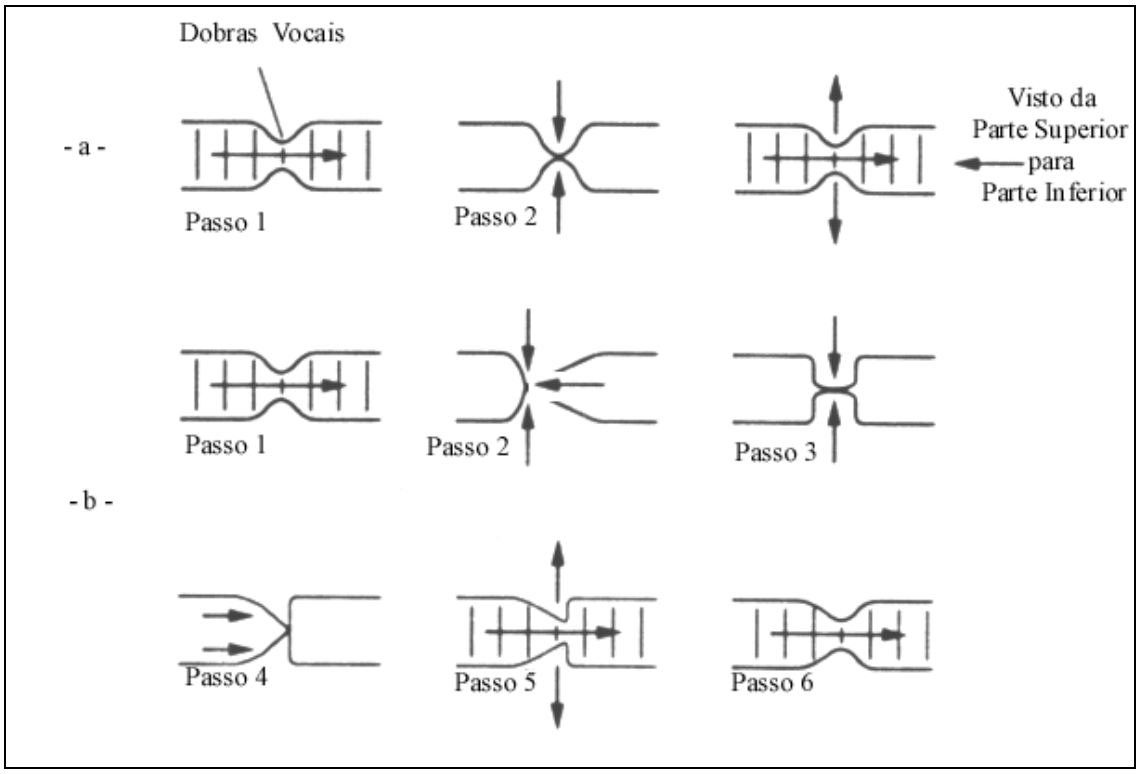


Figura 2.4 – Principio de Bernoulli aplicado à glote: (a) basicamente horizontal abertura/fechamento do ciclo da voz; (b) com movimento vertical das dobras. Linhas verticais representam o fluxo de ar na direção das setas [22].

Suponha inicialmente que as pregas estejam soltas e abertas. A contração dos pulmões provoca inicialmente um fluxo de ar pela glote, conforme a propriedade dinâmica dos fluidos chamada *Principio de Bernoulli* como a velocidade da corrente de ar (isto é, a velocidade das partículas de ar), diminuindo de fato a pressão local na região da glote. Ao mesmo tempo, a tensão das pregas, juntamente com a queda da pressão da glote causam o fechamento abruptamente das pregas vocais.A pressão criada entre elas e os pulmões em continua contração, forçam a sua abertura. O processo inteiro então se repete, e os resultados são, saídas periódicas de ar que entraram no trato vocal.

Assim, posteriormente ilustra-se as pregas vocais como oscilações horizontais, perpendiculares da parede da traquéia. O movimento dessas pregas, porém, não é muito simples. Por exemplo, ambos movimentos horizontais e verticais das pregas podem ocorrer simultaneamente, como ilustra na Figura 2.4 (b) durante a fase em que a glote está aberta,

isso deve-se ao fato que as partes superiores são mais flexíveis que as inferiores. Há uma certa demora no fechamento das duas regiões como mostra a Figura 2.4(b) nos passos 1-3. O movimento adicional vertical existe devido a esta demora entre a abertura das duas regiões. Quando a pressão de ar aumenta, durante o tempo em que a glote está fechada, a região inferior das dobras é a primeira a ser empurrada para cima, seguida pela região superior, como visto nos passos 4-6. Tal complexidade conduziu a um modelo não linear para massa dupla [12] mostrado na Figura 2.5, bem como a um modelo não-linear multicomponente que descreve os vários modos vibrantes ao longo das próprias dobras [28]. As massas m_k, constante de alimentação não-linear s_k, e constante de lubrificação τ_k no qual correspondem ao modelo mecânico respectivamente da massa, tensão e resistência interna das dobras vocais e da cartilagem externa.

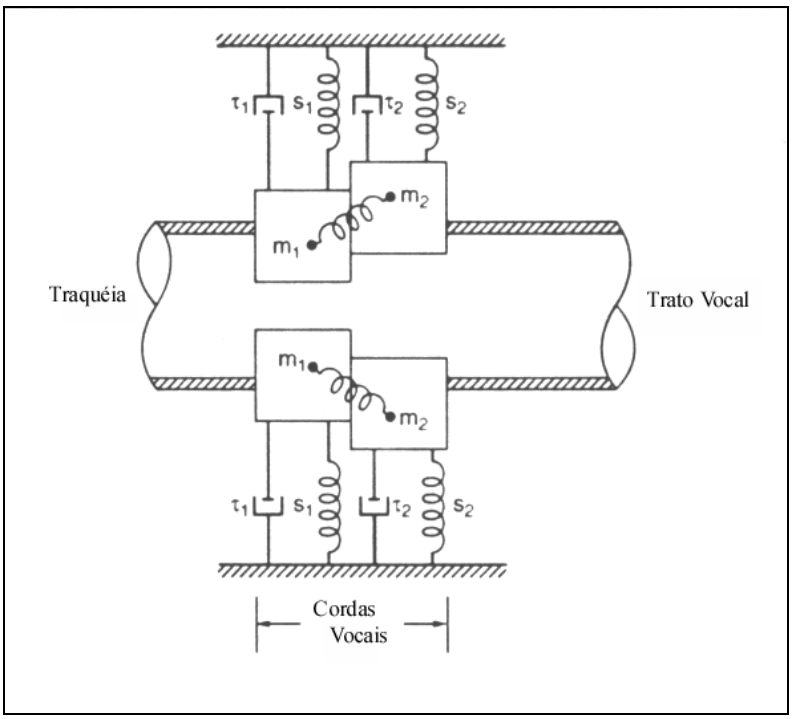


Figura 2.5 – Modelo mecânico de massa dupla de Flanagan e Ishizaka, com massas m_1 e m_2, resistências τ_1 e τ_2, e fontes constantes s_1 e s_2 [13].

De acordo com a descrição da velocidade do fluxo de ar na glote, se medirmos a velocidade do fluxo de ar da glote em função do tempo obtém-se aproximadamente uma forma de onda similar a ilustrada na Figura 2.6, que não segue exatamente a área tempo-variante da glote. Tipicamente, com as pregas fechadas, o fluxo inicia-se lentamente, até atingir o máximo, e rapidamente cai a zero quando as dobras vocais fecham-se abruptamente. No intervalo de tempo em que elas estão fechadas, não existe fluxo de ar. Esse intervalo refere-se como a fase fechada da glote. O intervalo de tempo em que não é zero o fluxo de ar e que aumenta até atingir o fluxo máximo de ar é chamado de fase aberta da glote. O tempo entre o fluxo máximo e o fechamento da glote é referido com fase de retorno. A forma específica do fluxo de ar pode mudar de acordo com o locutor, forma de falar e o tipo de som emitido. Em alguns casos, as dobras vocais não se fecham completamente, não existindo assim a fase fechada. Com o objetivo de simplificar, ao longo deste capítulo trataremos a velocidade do fluxo de ar da glote apenas como fluxo glotal.

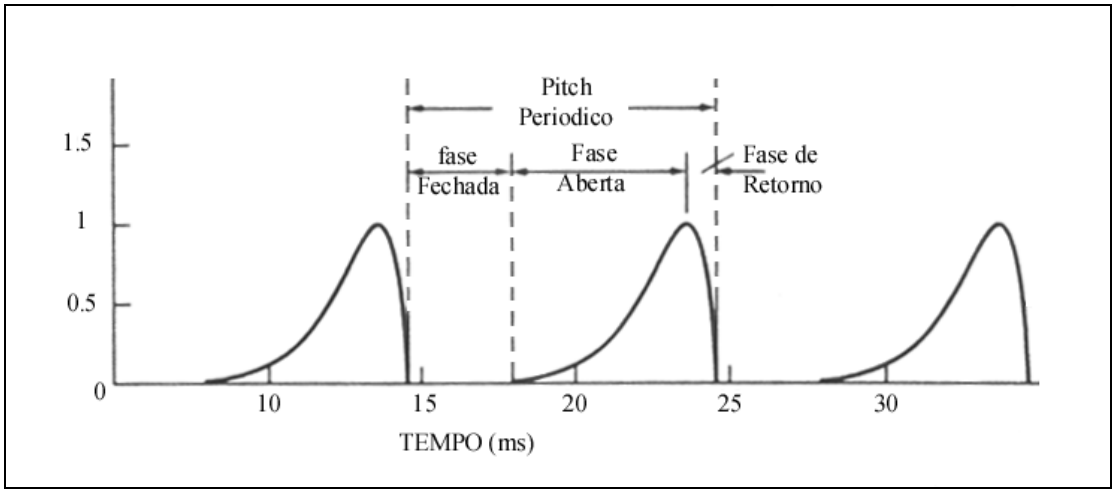


Figura 2.6 – Ilustração do período da velocidade de fluxo glotal.

A duração de um ciclo glotal é referido como período do pitch e reciprocamente o período de pitch corresponde ao pitch, alguns [22] referem-se como freqüência

fundamental. O termo pitch pode causar confusão porque é usado freqüentemente para descrever subjetivamente a percepção de altura de um complexo som musical até mesmo quando não é única a freqüência fundamental. Nesta dissertação, usa-se estritamente o termo pitch para referir à freqüência fundamental. Em conversas, durante os sons das vogais, nós pode-se encontrar de um a quatro períodos de pitch durante a emissão do som, embora o número do período de pitch mude com numerosos fatores como tensão e taxa de fala. A taxa com que as pregas oscilam entre a fase aberta, fechada e ciclo de retorno é influenciada por muitos fatores. Nestes incluem a tensão muscular das pregas (como aumento da tensão, ocorrendo o mesmo com o pitch), a massa das pregas vocais (aumento da massa, queda do pitch porque as pregas estão mais lentas), e a pressão de ar entre a glotes na laringe e na traquéia, o que poderá diminuir em um som acentuado ou em um maior estágio de excitação da fala (como a pressão cai a glote diminui, ocorrendo o mesmo com o pitch). O pitch varia de 60 Hz a 400 Hz. Comumente, os homens possuem um pitch mais baixo que as mulheres, visto que as dobras vocais são maiores e possuem mais massa.

2.2.3. O trato vocal

O trato vocal é composto da cavidade oral que segue da laringe até a boca e uma passagem nasal que é unida a ele pelo velum. O trato oral é encontrado em muitas dimensões diferentes e a seção transversal pelo movimento da língua, dentes, lábios, mandíbula possui um comprimento médio de 17 cm em um adulto do sexo masculino sendo menor no sexo feminino, e a variação espacial da seção transversal de até 20 cm^2. Se fosse ouvida a pressão da onda, externo às dobras vocais durante o processo de fala, ouveria-se apenas um zumbido o que não seria muito interessante. Uma proposta inicial do

trato vocal é “modelar” a fonte, o que é importante na percepção distinta dos sons da fala. Uma segunda proposta é a geração de novas fontes de sons.

A modulação espectral, segue algumas condições, a relação entre a velocidade na entrada do fluxo de ar glotal e na saída fluxo de ar do trato vocal podem ser aproximados por um filtro linear de ar com ressonância, muito usado em órgão de transporte e aparelhos de ar. As freqüências ressonantes do trato vocal são no contexto cientifico de voz chamado freqüência formante ou simplesmente formante. A palavra “formante” referisse a contribuição total da ressonância. Assim usa-se para expressar “largura de banda do formante” e “amplitude do formante” o termo freqüência do formante. Os formantes mudam com diferentes constituições do trato vocal. Por exemplo, com diferentes vogais, a mandíbula, os dentes, os lábios e a língua assumem posições diferentes. A letra (a) da Figura 2.7 ilustra este fato.

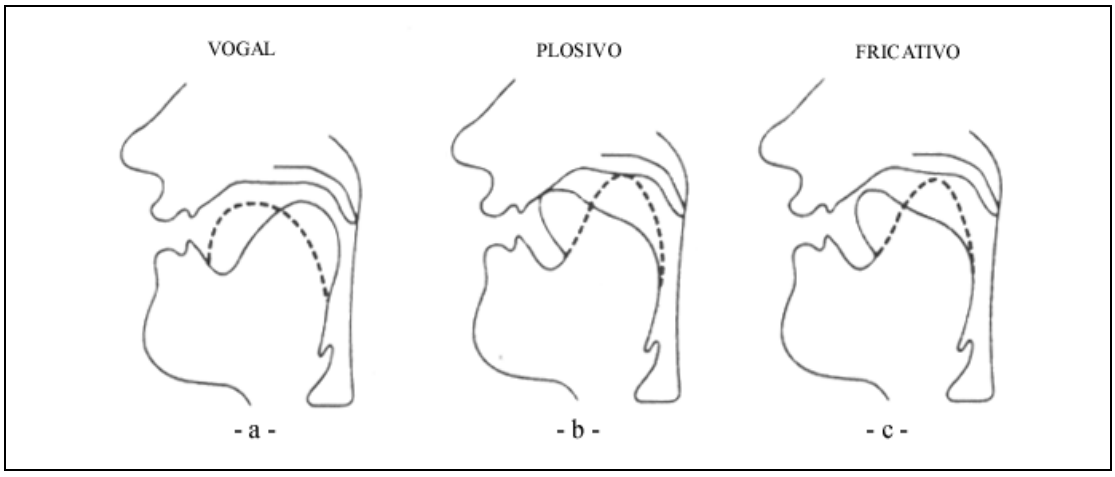


Figura 2.7 – Mudanças ocorridas no trato vocal para as formas de: (a) vogais (com alimentação periódica), (b) plosivos (com alimentação impulsiva) e (c) fricativos (com alimentação ruidosa) [22].

Os picos do espectro da resposta do trato vocal correspondem aproximadamente a seus formantes. Mais especificamente, quando esse é modelado como um sistema linear

invariante no tempo “todo-pólo”, um pólo em $z_0 = r_0 e^{j\omega_0}$ corresponde aproximadamente ao formante do trato vocal. A freqüência do formante é $\omega = \omega_0$ e a largura de faixa dele é determinada pela distância do pólo ao circulo unitário (r_0). Pelo fato dos pólos serem tipicamente uma seqüência real, os pólos complexos aparecem em pares (exceto quando estão no eixo real). Somente freqüências positivas são usadas na definição da freqüência do formante e a largura de faixa do formante também é obtido utilizando-se as freqüência positivas.

Foi visto que as formas de onda correspondem a diferentes cavidades ressonantes; diferentes formas do trato vogal podem resultar também em diferentes fontes de sons. A Figura 2.7 (b) mostra o fechamento completo do trato vocal, a língua pressionando novamente o palato, provocando uma fonte impulsiva de som. Existe uma formação da pressão entre o fechamento e então liberação abrupta da pressão. A Figura 2.7 (c) mostra a formação de outro som criado com a língua no palato, mas não completamente impedido, para a geração da turbulência e a formação dos ruídos.

Ainda existe um outro tipo fonte que é gerado no trato vocal, mas é menos compreendida que as fontes ruidosas e impulsivas ocorridas na constrição do trato oral. Essas fontes surgem da interação dos vórtices com os limites do trato vocal assim como o trato vocal falso, dentes ou oclusões do trato oral [3, 30]. Uma noção sobre a natureza de um vórtice; para o momento, basta que seja pensado em vórtice na área oral como uma corrente de ar de minúscula rotação. Durante a fala, os vórtices movem-se possivelmente como um trem de glote para os lábios ao longo da área oral para a fala, e são preditos a iniciar o jato de ar que emana da glote durante a vibração da prega vocal [3, 30]. Vórtices também podem surgir durante sons fricativos com fontes resultantes distribuídas ao longo da área oral [16]. Existem tendências da influência temporal e espectral dos vórtices nas fontes, e talvez em parte nas características dos sons da fala [3, 16, 30].

2.2.4. Classificação do som pela fonte

Existem várias formas de classificar um som. Por exemplo, quando classificam-se os sons baseados nas diferentes fontes do trato vocal, foi visto que aquelas fontes diferentes estão relacionadas ao estado da prega vocal, mas são também formadas várias construções no trato oral. Os sons da fala gerados com uma fonte glotal periódica são denominados *"voiced"*, analogamente, os sons que não são gerados dessa forma são chamados de *"unvoiced"*. Existe uma variedade de sons *"unvoiced"*, incluindo esses criados com fontes ruidosas da constrição do trato oral. Pelo fato de tais tipos de sons serem formados pela fricção do movimento ar, contrário a constrição, esses sons são conhecidos por fricativos, como mostrado na Figura 2.7 (c).

2.3- Propagação do Som [8]

Serão apresentados alguns princípios físicos envolvidos no processo de geração da fala com a intenção de obter o equacionamento matemático do mesmo.

Som é geralmente sinônimo de vibração. As ondas sonoras são produzidas por vibrações e propagadas no ar, ou outro meio, através das vibrações das partículas do mesmo.

As leis da física são as bases para descrever a geração e a propagação dos sons no sistema vocal. Em particular as leis fundamentais de conservação de massa, momento e energia, em conjunto com as leis da termodinâmica e da mecânica de fluidos, são aplicáveis aos sons da fala. Um conjunto de equações diferenciais e parciais é obtido a partir desses princípios que retratam o movimento do ar pelo sistema vocal. A solução dessas equações é extremamente difícil, a menos que sejam feitas algumas suposições

simplificadoras sobre a forma do trato vocal e às perdas de energia do sistema. É o que os modelos propostos tentam fazer. Um modelo ideal segundo a teoria acústica deveria considerar os seguintes efeitos:

1. variação da forma do trato vocal no tempo;
2. perdas devido a condução do calor e atrito viscoso nas paredes do trato vocal;
3. suavidade das paredes do trato vocal;
4. radiação do som nos lábios;
5. acoplamento nasal; e;
6. excitação do som no trato vocal.

Entretanto é impraticável considerar todos esses fatores para uma análise matemática.

Uma configuração física simplificada e útil na interpretação do processo de produção da fala é representada na Figura 2.8. O trato vocal é modelado como um tubo de seção transversal não uniforme e variante no tempo. Não são consideradas as perdas de energia no volume do fluido ou nas paredes do tubo devido a condução do calor ou viscosidade.

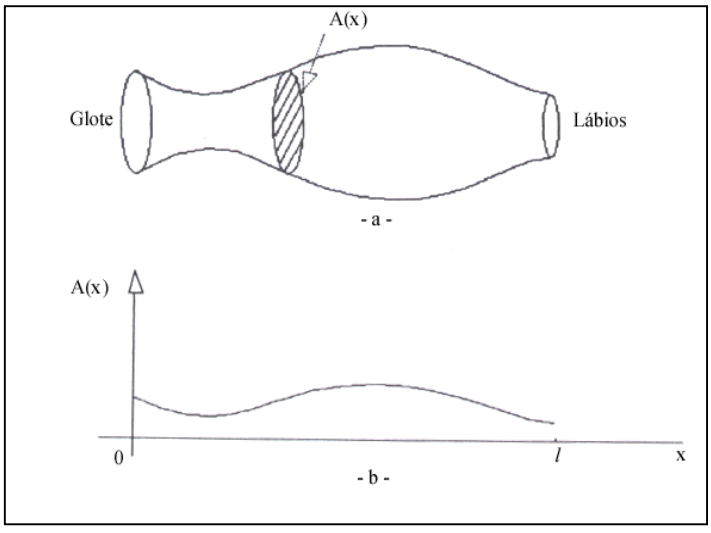


Figura 2.8 – (a) Diagrama simplificado do trato vocal; (b) Função área correspondente. [11].

Com estas suposições, e usando as leis de conservação da massa, do momento e da energia, Portnoff [32] mostrou que as ondas sonoras neste tubo satisfazem o seguinte par de equações:

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u / A)}{\partial t} \tag{2.1a}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2} \frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t} \tag{2.1b}$$

onde

$p = p(x,t)$ é a variação da pressão sonora no tubo na posição x e no tempo t.

$u = u(x,t)$ é a variação do fluxo do tempo de escoamento na posição x e no tempo t.

ρ é a densidade do ar no tubo.

c é a velocidade do som.

$A = A(x,t)$ é a "função área" do tubo.

Um conjunto de equações similares foi desenvolvido também por Sondhi [23].

A solução completa para este sistema não é possível a não ser para a configuração mais simplificada. Onde uma solução numérica pode ser obtida. A solução completa das equações diferenciais requer que a pressão p e o fluxo de escoamento sejam conhecidos em cada x na região limitada entre a glote e os lábios. As condições de contorno nos extremos devem considerar o efeito da radiação do som nos lábios e os efeitos pela natureza da excitação na glote (ou em algum ponto interno). Em adição às condições de contorno, a função área do trato vocal $A(x,t)$ deve ser conhecida. A Figura.2.8 mostra a função área do tubo em um determinado tempo.

2.4 - Tubo Uniforme Sem Perdas [8]

Algum discernimento sobre a natureza do sinal da fala pode ser obtido considerando um modelo bem simplificado o que torna possível a solução das Equações (2.1). Este modelo onde a função área do trato vocal é considerada constante em *x* e *t* (invariante no tempo e de seção transversal uniforme) é composto por um tubo de seção constante excitado por uma fonte ideal de fluxo de escoamento.Ele pode ser representado por um pistão que causa movimento em qualquer sentido, independente de variações de pressão no tubo, como mostra na Figura 2.9. Este modelo apesar de distante do modelo real, pode ser como uma aproximação básica para a análise das características essenciais que resultam da solução, em comum com os modelos mais realistas. Modelos mais gerais podem ser construídos pela concatenação de tubos uniformes.

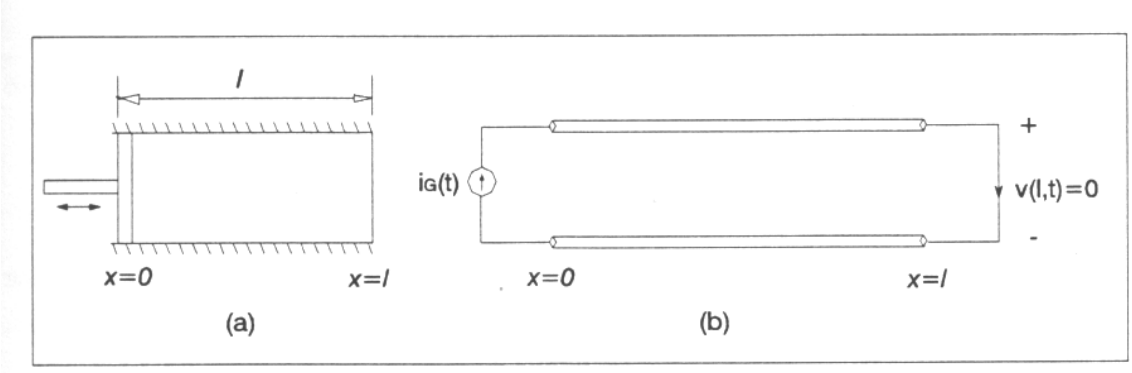


Figura 2.9 – (a) Tubo uniforme sem perdas; (b) Linha de transmissão elétrica análoga. [11]

Se *A(x,t)* = *A* é constante, então as Equações (2.1) resumem para a forma:

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t} \tag{2.2a}$$

$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t} \tag{2.2b}$$

As soluções para as Equações (2.2) são da forma

$$u(x,t) = \left[u^{+}\left(t - \frac{x}{c} \right) - u^{-}\left(t + \frac{x}{c} \right) \right] \tag{2.3a}$$

$$p(x,t) = \frac{\rho c}{A}\left[u^{+}\left(t - \frac{x}{c} \right) + u^{-}\left(t + \frac{x}{c} \right) \right] \tag{2.3b}$$

As funções $u^{+}\ (t - x/c)$ e $u^{-}\ (t + x/c)$ representam ondas viajantes nas direções positiva e negativa, respectivamente. As relações entre elas são determinadas pelas condições de contorno.

Na teoria das linhas de transmissão elétrica, uma linha uniforme sem perdas, a voltagem $v(x,t)$ e a corrente $i(x,t)$ na linha de transmissão estão relacionadas pelas equações:

$$-\frac{\partial v}{\partial x} = L\frac{\partial i}{\partial t} \tag{2.4a}$$

$$-\frac{\partial i}{\partial x} = C\frac{\partial v}{\partial t} \tag{2.4b}$$

onde:

$v = v(x,t)$ é a voltagem

$i = i(x,t)$ é a corrente

L é a indutância por unidade de comprimento

C é capacitância por unidade de comprimento

Comparando as Equações (2.2) e (2.4) verifica-se que são análogas, do mesmo tipo. Sendo assim a teoria de linhas de transmissão elétrica também se aplica aos tubos uniformes, desde que se faça a analogia contida na Tabela (2.1).

Tabela 2.1 – Analogia entre tubos acústicos uniformes e linhas de transmissão elétrica.

Quantidade Acústicas	**Quantidades Elétricas**
p – pressão	v – voltagem
u – escoamento	i – corrente
ρ / A – indutância acústica	L – indutância
$A/(\rho c^2)$ – capacitância acústica	C – capacitância

O tubo comporta-se como uma linha de transmissão terminado com um curto circuito ($v(1,t) = 0$) e no início excitado por uma fonte de corrente $(i(0,t) = i_G(t))$, com mostra a Figura 2.9.

A representação no domínio da freqüência de sistemas lineares, como linha de transmissão, é bastante útil. Por analogia pode-se obter representação similar para o tubo acústico sem perdas. Assumindo a condição de contorno em $x = 0$

$$u(0,t) = u_G(t) = U_G(\Omega)e^{j\Omega t} \tag{2.5}$$

ou seja, o tubo é excitado por uma variação exponencial complexa da variação de escoamento de freqüência radial Ω e amplitude complexa $U_G(\Omega)$. Como as equações são lineares, então:

$$u^+(t - x/c) = K^+ e^{j\Omega(t-x/c)} \tag{2.6a}$$

$$u^{-}(t + x/c) = K^{-} e^{j\Omega(t+x/c)} \tag{2.6b}$$

Substituindo as Equações (2.6) nas Equações (2.3) e aplicando-se as condições de contorno:

$$p(l,t) = 0 \tag{2.7}$$

na extremidade do tubo correspondente aos lábios, e a equação (2.5) na outra extremidade obtem-se os valores das constantes K^{+} e K^{-}:

$$K^{+} = \frac{U_G(\Omega)}{1 + e^{-2j\frac{\Omega l}{c}}} \tag{2.8a}$$

$$K^{-} = -\frac{U_G(\Omega)}{1 + e^{2j\frac{\Omega l}{c}}} \tag{2.8b}$$

Substituindo-se K^{+} e K^{-} nas Equações (2.6) e estas nas Equações (2.3) obtém-se então às soluções senoidais em regime permanente para *p(x,t)* e *u(x,t).*

$$p(x,t) = jZ_0 \frac{\text{sen}[\Omega(l-x)/c]}{\cos[\Omega l/c]} U_G(\Omega) e^{j\Omega t} \tag{2.9a}$$

$$u(x,t) = \frac{\cos[\Omega(l-x)/c]}{\cos[\Omega l/c]} U_G(\Omega) e^{j\Omega t} \tag{2.9b}$$

onde:

$Z_0 = \frac{\rho c}{A}$ é por analogia, chamado impedância acústica característica do tubo.

Expressando *p(x,t)* e *u (x,t)* diretamente por uma excitação complexa, evita-se soluções com ondas viajantes:

$$p(x,t) = P(x,\Omega)e^{j\Omega t} \tag{2.10a}$$

$$u(x,t) = U(x,\Omega)e^{j\Omega t} \tag{2.10b}$$

Substituindo as Equações (2.10) na Equação (2.1) obtém-se as Equações relacionadas às amplitudes complexas:

$$-\frac{dP}{dx} = ZU \tag{2.11a}$$

$$-\frac{dU}{dx} = YP \tag{2.11b}$$

onde

$Z = j\Omega\frac{\rho}{A}$ é a impedância acústica por unidade de comprimento

$Y = j\Omega\frac{A}{\rho c^2}$ é a admitância acústica por unidade de comprimento.

As Equações (2.11) têm soluções da seguinte forma:

$$P(x,\Omega) = Ae^{yx} + Be^{-yx} \tag{2.12a}$$

$$U(x,\Omega) = Ce^{yx} + De^{-yx} \tag{2.12b}$$

onde

$$y = \sqrt{ZY} = j\Omega / c \tag{2.13}$$

Os coeficientes desconhecidos são calculados aplicando as condições iniciais

$$P(1,\Omega) = 0 \tag{2.14b}$$

$$U(0,\Omega) = U_G(\Omega) \tag{2.14b}$$

O resultado será o mesmo encontrado na Equação (2.9). As Equações (2.9) expressam a relação entre a fonte de escoamento senoidal, a pressão e o escoamento em qualquer ponto do tubo. Calcula-se a razão entre a vazão nos lábios e na fonte de escoamento, obtém-se:

$$u(l,t) = U(l,\Omega)e^{j\Omega t} = \frac{1}{\cos(\Omega l / c)} U_G(\Omega)e^{j\Omega t} \tag{2.15}$$

A razão que segue é então a resposta em freqüência do sistema que relaciona vazão de entrada e saída,

$$\frac{U(l,t)}{U_G(\Omega)} = V_\alpha(j\Omega) = \frac{1}{\cos(\Omega l / c)} \tag{2.16}$$

Esta função é plotada na Figura. 2.10(a) para os valores 1 = 17,5 cm e c = 35000 cm/seg.

Substituindo-se Ω por *s/j*, obtém-se a transformada de Laplace ou função do sistema:

$$V_\alpha(s) = \frac{2e^{-sl/c}}{1+e^{-s2l/c}} \tag{2.17}$$

Nota-se que $V_\alpha(s)$ tem um número infinito de pólos igualmente espaçados sobre o eixo $j\Omega$ na Figura 2.10 (b):

$$s_n = \pm j\left[\frac{(2n+1)\pi c}{2l}\right] \qquad n = 0,\pm 1,\pm 2,... \qquad (2.18)$$

Os pólos de uma função do sistema, para um sistema linear no tempo correspondem às freqüências naturais do sistema, ou ainda, as freqüências de ressonância. Considerando a produção da fala, tais freqüências são as já citadas neste capítulo, *freqüências formantes*.

Recordamos então que a função resposta em freqüência permite determinar a resposta do sistema não somente para funções senoidais, mas para qualquer entrada arbitrária através da análise de Fourier. Assim a resposta em freqüência é uma conveniente caracterização para o modelo do sistema vocal.

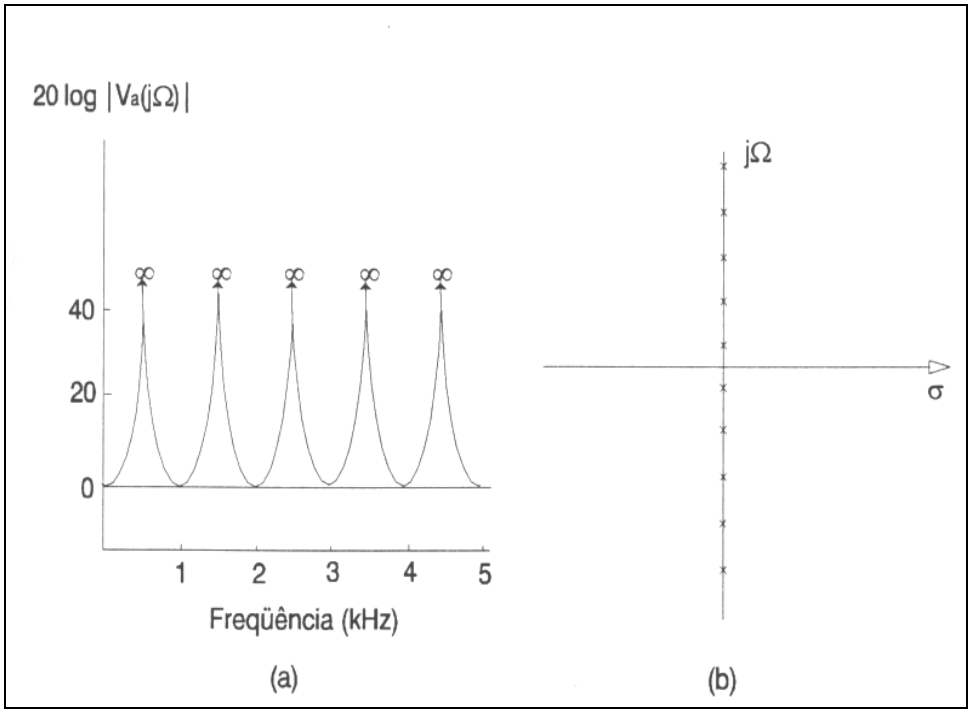


Figura 2.10 – (a) Resposta em freqüência; (b) Localização dos pólos para um tubo uniforme sem perdas [11].

2.5 - Modelo dos Tubos Uniformes Concatenados Sem Perdas [8]

O modelo útil na produção da fala é baseado na superposição de que o trato vocal pode ser representado como uma concatenação de tubos acústicos sem perdas, como representado na Figura 2.11.

As áreas constantes A_k dos tubos são escolhidas de forma a aproximarem da área *A(x)* do trato vocal. Se uma grande quantidade de tubos de pequenos comprimentos é usada, é razoável esperar que as freqüências de ressonância dos tubos tendem a se aproximarem das freqüências de ressonância de um único tubo com área variando continuamente. Entretanto esta aproximação despreza as perdas devido à condução de calor, atrito, vibração das paredes e outras. Portanto, apesar da aproximação em formato, esse modelo continua sendo diferente do modelo detalhado que inclui as perdas.

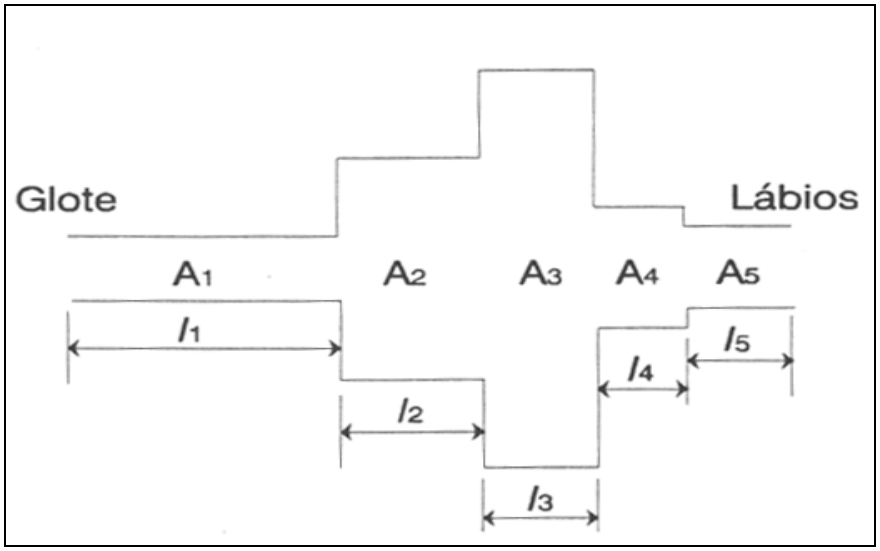


FIGURA 2.11 – Concatenação de 5 tubos acústicos uniformes sem perdas [11].

O mais importante desta discussão é o fato de que o modelo de tubos sem perdas representa uma conveniente transição entre os modelos contínuos e os modelos discretos no tempo.

2.6 - Modelos Digitais da Fala

A acústica da produção da fala fornece um modelo matemático detalhado. A teoria acústica relaciona as características da fala com a física da produção da mesma.

Os sons da fala são gerados em três modos e cada um resulta em um tipo distinto de saída. O aparelho vogal impõe em suas ressonâncias sobre a excitação para produzir os diferentes tipos de sons da fala.

Na Figura 2.12 pode-se verificar o modelo de um sistema linear que produz uma saída com as propriedades desejadas quando controlado por um conjunto de parâmetros relacionados com o processo de produção dos sons da fala. Apesar de ser um modelo equivalente ao modelo físico em sua saída a estrutura interna não imita a física da produção da fala.

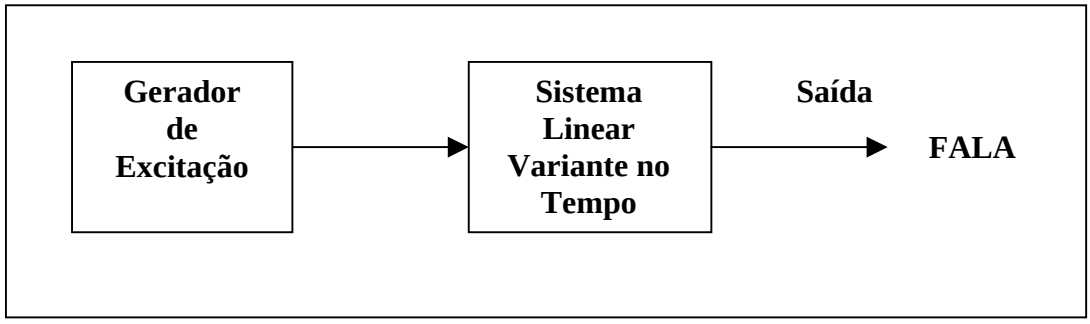


Figura 2.12 – Modelo esquematizado da produção da fala, “Modelo terminal analógico” [11].

Para produzir um sinal da fala, o modo de excitação e as propriedades de ressonância do sistema linear precisam variar com o tempo. As propriedades do sinal da fala mudam lentamente com o tempo. Em muitos sons da fala é razoável supor que as propriedades gerais da excitação e do trato vocal permaneçam fixas entre 10 e 20 *ms*.

Assim, um modelo terminal analógico envolve um sistema linear variante no tempo excitado por um sinal cuja natureza básica muda de pulsos quase periódicos (em sinais sonoros) para ruídos aleatórios (em sinais surdos).

As características mais relevantes do modelo discreto no tempo de tubos sem perdas são representadas na Figura 2.13, que traz o trato vocal caracterizado por um conjunto de áreas, ou equivalentemente, pelos coeficientes de reflexão. A relação entre a entrada e saída representada por uma função de transferência, V(z) é obtida por:

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}} \tag{2.19}$$

onde G e $\{\alpha_k\}$ dependem da função área. Qualquer sistema tendo a função de transferência da Equação (2.19) produz a mesma saída, em resposta a uma determinada entrada. Assim, modelos terminais analógicos tomam a forma geral da Figura 2.13 (b).

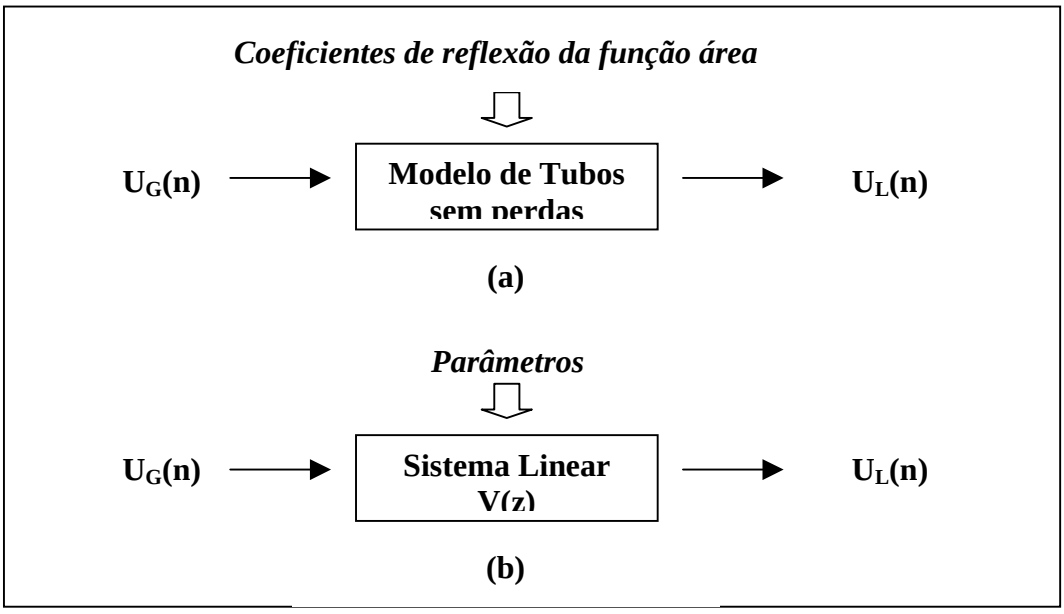


FIGURA 2.13 – (a) Diagrama de blocos representando o modelo de tubos sem perdas; (b) Modelo terminal analógico [11].

Para obter um sistema mais completo deve-se incluir uma função de mudança de excitação e os efeitos da radiação do som nos lábios à resposta do trato vocal.

No trato vocal as ressonâncias (formantes) da fala correspondem aos pólos da função V(z). Um modelo todo-polo é uma boa representação do aparelho vocal para a maioria dos sons; entretanto, a teoria acústica diz que sons nasais e fricativos requerem tanto ressonância quanto anti-ressonância (pólos e zeros). Nesses casos pode-se incluir zeros na função transferência.

Se os coeficientes do denominador V(z) são reais, as raízes do polinômio do denominador são também reais ou ocorrerão em pares complexos conjugados. Uma freqüência de ressonância típica do trato vocal é obtida por:

$$S_k, S_k^* = -\alpha_k \pm j2\pi f_k \tag{2.20}$$

Os correspondentes pólos complexos conjugados na representação de tempo discreto são:

$$z_k, z_k^* = e^{-\sigma_k T} e^{\pm j 2\pi F_k T} = e^{-\sigma_k T}\cos(2\pi F_k T) \pm j e^{-\sigma_k T}\,\mathrm{sen}(2\pi F_k T) \tag{2.21}$$

A largura de banda da ressonância do trato vocal é aproximadamente $2\alpha_k$ e a freqüência central é $2\pi F_k$. No plano *z* o raio da origem para o pólo determina a largura de banda, ou seja:

$$|z_k| = e^{-\sigma_k T} \tag{2.22a}$$

e o ângulo no plano *z* é:

$$\theta_k = 2\pi F_k T \tag{2.22b}$$

Assim, se o denominador de *V(z)* for fatorado, as correspondentes freqüências análogas formantes e a largura de banda podem ser encontradas usando as Equações (2.22). Como mostra a Figura 2.14, as freqüências complexas naturais do trato vocal humano estão todas na metade do plano *s* desde que ele seja um sistema estável. Então, $\alpha_k > 0$, portanto $|z_k| < 1$, isto é, todos os pólos correspondentes do modelo do tempo discreto precisam estar dentro do circulo unitário como é requerido para a estabilidade. A Figura 2.14 representa as freqüências ressonantes complexas típicas em ambos os planos s e z.

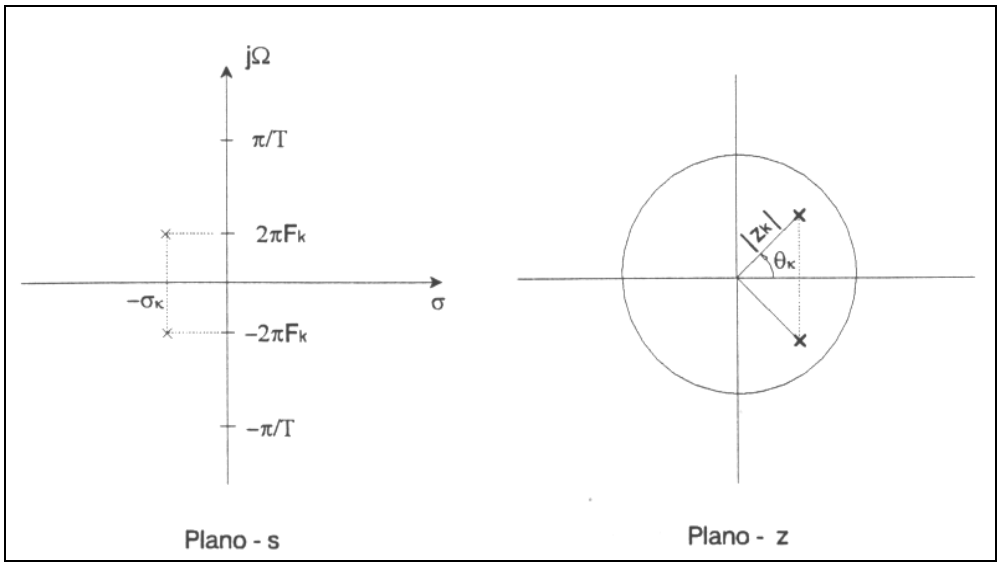


Figura 2.14 – Representações das ressonâncias do trato vocal: (a) Plano *s*; (b) Plano *z*.

Para o efeito de radiação nos lábios um aproximação razoável é obtida com:

$$R(z) = R_0(1 - z^{-1}) \tag{2.23}$$

A Figura 2.15 representa um modelo incluindo o trato vocal e os efeitos da radiação nos lábios.

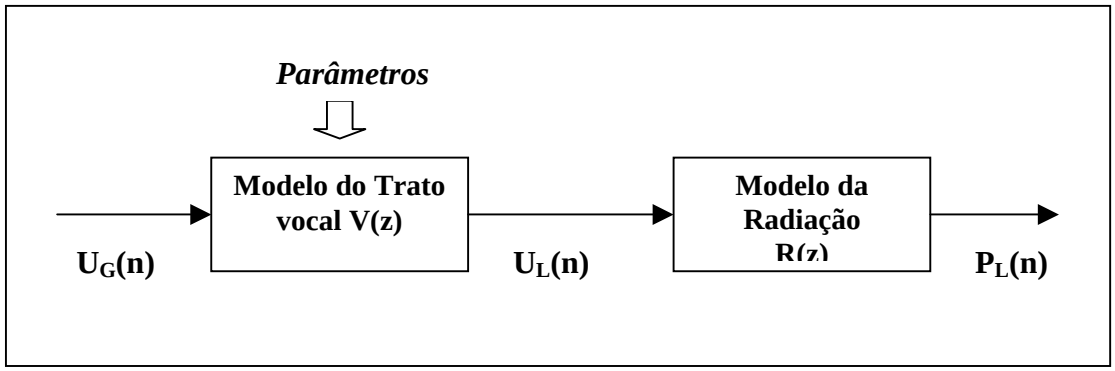


Figura 2.15 – Modelo terminal analógico incluindo os efeitos da radiação nos lábios.

Para complementar o modelo terminal analógico, é necessário gerar um sinal de entrada (sinal de excitação) apropriado para o sistema do trato vocal. Relembrando que a maioria dos sons da fala podem ser classificados como sonoros ou surdos, pode-se ver em termos gerais que é requerido uma fonte que produza ondas em formas de pulsos quase periódicos ou em forma de ruído aleatório.

A Figura 2.16 esquematiza a geração do sinal de excitação para sons sonoros, isto é, a onda glotal. O gerador de trem de impulsos produz uma seqüência de impulsos unitários espaçados por um período fundamental (período de pitch) desejado. Este sinal excita um sistema linear cuja resposta impulsiva *g(n)* tem a forma de onda glotal desejada. O controle de ganho, A_v, controla a intensidade da excitação sonora.

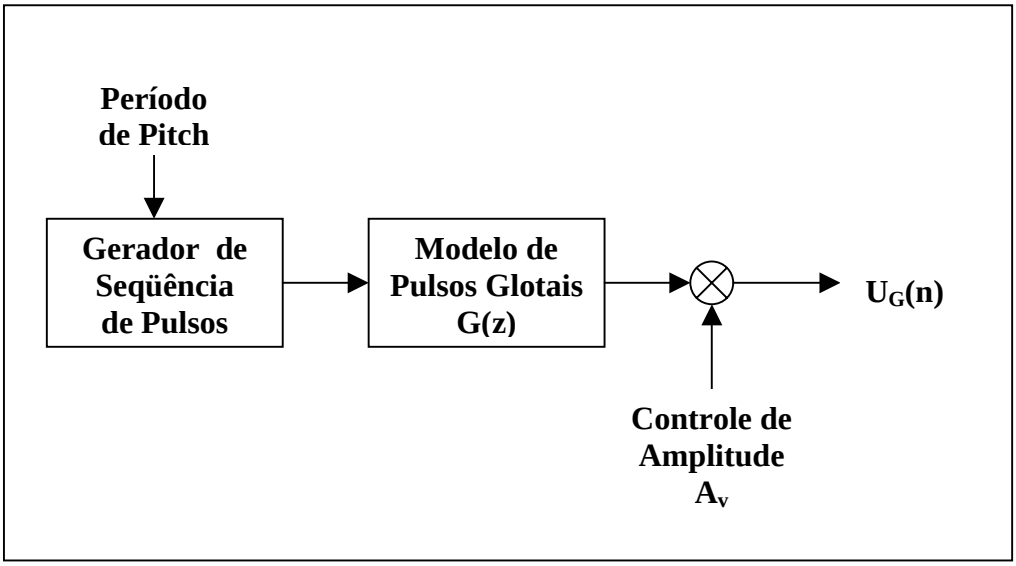


Figura 2.16 – Geração do sinal de excitação para sons sonoros [11].

Rosenberg [24], em estudos do efeito da forma do pulso glotal na qualidade da fala, descobriu que a forma de onda glotal natural poderia ser substituída por uma forma de onde de pulso sintético como:

$$g(n) = \begin{cases} \frac{1}{2}\left[1 - \cos\left(\frac{\pi n}{N_1}\right)\right] & 0 \le n \le N_1 \\ \cos\left(\frac{\pi(n - N_1)}{2N_2}\right) & N_1 \le n \le N_1 + N_2 \\ 0 & \text{outros casos} \end{cases} \qquad (2.24)$$

A transformada *z* de *g(n)*, *G(z)*, tem apenas zeros. Ela apresenta bons resultados como modelo de 2 pólos para *G(z)*.

Para sons surdos o modelo de excitação é muito mais simples. Tudo que é requerido é uma fonte de ruído aleatório e um parâmetro ganho, A_N, para controlar a intensidade da excitação surda. Para modelos de tempo discreto, um gerador de ruído aleatório produz uma fonte de ruído de espectro plano.

Um modelo completo com a junção dos elementos já descritos neste capítulo pode ser visto na Figura 2.17 em que uma chave entre o gerador de sons sonoros e surdos permite a escolha do modo de excitação.

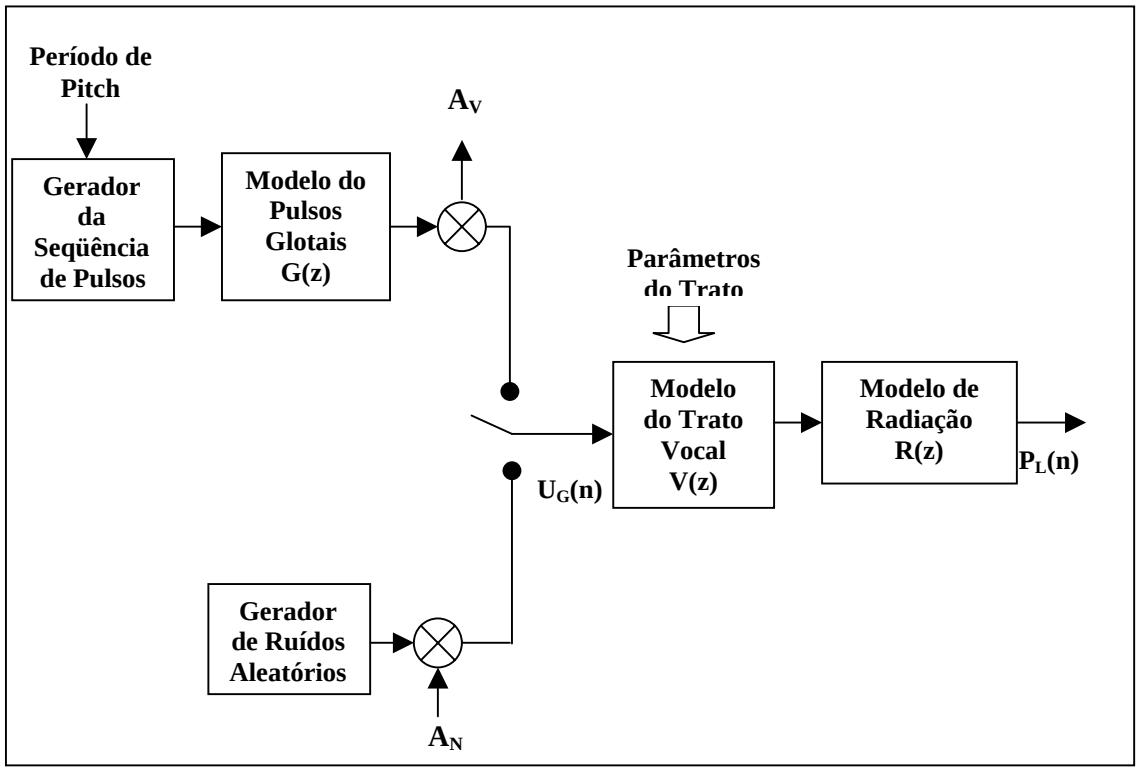


FIGURA 2.17 – Modelo geral para geração da fala em tempo discreto [11].

O trato vocal pode ser modelado de várias maneiras diferentes. Em alguns casos é conveniente agrupar os modelos do pulso glotal e a radiação em um único sistema. É conveniente combinar as componentes do pulso glotal, radiação e trato vocal todos juntos e representá-los como uma única função de transferência do tipo todo-polo,

$$H(z) = G(z).V(z).R(z) \tag{2.25}$$

Neste ponto, uma questão natural são as limitações deste modelo. As deficiências dele não limitam severamente sua aplicabilidade. Primeiro, existe a questão da variação dos parâmetros com o tempo. Em sons contínuos, tais como as vogais, os parâmetros mudam muito lentamente e o modelo trabalha muito bem. Com sons transientes tais como as consoantes oclusivas o modelo não é muito bom, mas ainda é

adequado. É necessário enfatizar que pode-se representar o sinal da fala em uma curta base de tempo. Os parâmetros do modelo são então assumidos constantes sobre os intervalos de tempo tipicamente de 10 a 20*ms*. A função de transferência *V(z)*, então, servem pra definir as estruturas de um modelo cujos parâmetros variam continuamente com o tempo. A segunda limitação é a falta de disposição de zeros como é requerido teoricamente para os sons nasais e fricativos. Essa é uma limitação para os sons nasais, mas não é muito severa para os sons fricativos. Zeros podem ser incluídos no modelo desejado. Terceira, a simples dicotomia da excitação sonoro-surda não é adequada para sons fricativos. Finalmente, o modelo requer que os pulsos sejam espaçados por um múltiplo inteiro do período de amostragem *T*. Winham e Steiglitz [23] têm considerado maneiras de eliminar esta limitação em situações onde é necessário um preciso controle de pitch.

2.7. Considerações Finais deste Capítulo

Este capítulo descreveu qualitativamente as principais funções no mecanismo da produção de fala e sua associação anatômica. Ao ser abordada a fisiologia e a anatomia pode-se ver como são formados os sons e porque as diferenças pequenas na constituição dos membros do aparelho vocal humano provocam grandes diferenças sonoras. Essas diferenças são as motivadoras do desenvolvimento de sistemas automáticos de reconhecimento do locutor pela voz.

Fica claro também a necessidade de um modelo físico matemático eficiente para representar a fala e a propagação sonora. Modelo este obtido da física acústica no processo de produção da fala que servirá de base para a aplicação das técnicas de processamento digital de sinais.

CAPÍTULO III

QUANTIZAÇÃO VETORIAL E MEDIDAS DE DISTÂNCIAS

3.1 - Introdução

O método de predição linear é uma técnica muito importante para estimar os parâmetros básicos de voz, como: freqüência fundamental, formantes, espectro e a função área do aparelho vocal.

Neste capítulo é feita análise LPC no domínio do tempo e no domínio da freqüência [18], [23]. Ele também trata das medidas de distorção. É mostrada a definição e são apresentadas as medidas mais usadas.

Este capítulo também trata das medidas de distorção. São dadas as definições e apresentadas as medidas mais usadas.

3.2 - Princípios Básicos de Análise de Preditor Linear

Este capítulo refere-se ao modelo de tempo discreto de produção de voz apresentado no capítulo II. Repete-se esse modelo fazendo algumas simplificações. A

forma particular desse modelo que é apropriada para a análise de predição linear é apresentada na Figura 3.1.

No modelo tudo-polo, assumi-se que o sinal de voz é uma combinação linear dos valores atrasados e a entrada u_n :

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{3.1}$$

Aplica-se a transformada z na Equação (3.1) obtém-se a função de transferência *H(z)*.

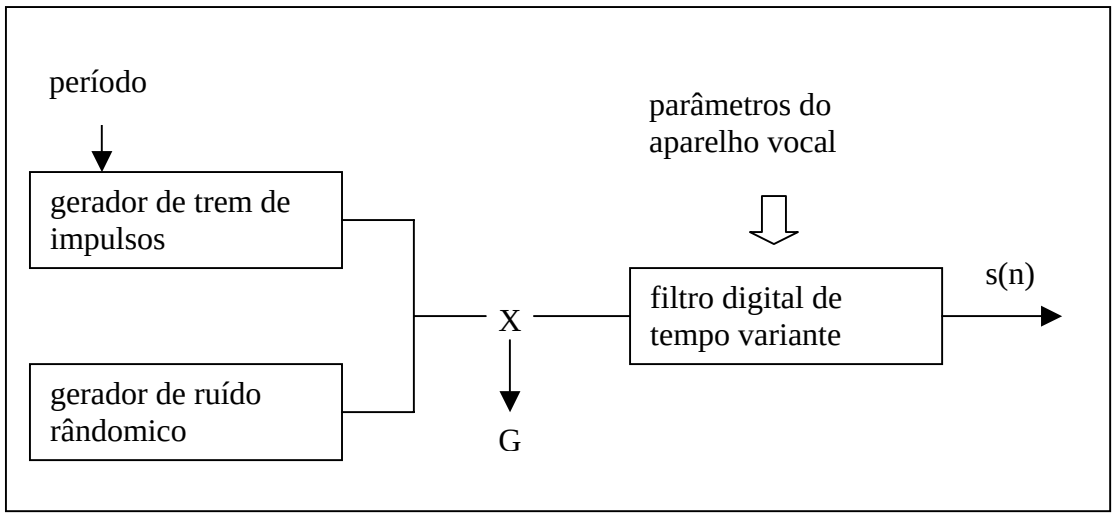


FIGURA 3.1- Diagrama de bloco simplificado para o modelo de produção de voz [23].

$$H(z) = \frac{G}{A(z)} \tag{3.2a}$$

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3.2b}$$

onde *G* é o fator ganho.

A função de transferência do filtro digital de variante no tempo apresentado na Figura 3.1 é *H(z)*. O modelo simplificado (tudo-pólo) é uma representação natural de sons nasais. Para sons nasais e fricativos, de acordo com a teoria acústica, devem-se incluir pólos e zeros na função de transferência. Mas se a ordem *p* é suficientemente alta, o modelo tudo-pólo produz uma boa representação para quase todos os sons do aparelho vocal. A maior vantagem deste modelo é que o parâmetro *G* e os coeficientes do filtro podem ser estimados de maneira direta e a computação é eficiente, pelo método de predição linear.

Precisa-se obter o ganho *G* e o conjunto de coeficientes do preditor a_k diretamente do sinal de voz, de maneira a obter uma boa estimação das propriedades espectrais do sinal de voz pela Equação (3.2). Os coeficientes do preditor precisam ser colocados em curtos intervalos de tempo devido a natureza variante no tempo do sinal de voz. A idéia básica é encontrar o conjunto de coeficientes do preditor que minimiza o erro médio quadrático sobre um curto segmento voz. Os parâmetros resultantes são os parâmetros do sistema, a função de transferência *H(z)*, do modelo de produção de voz.

Assume-se que a entrada um é desconhecida. Todavia, o sinal s_n pode ser predito aproximadamente por uma combinação linear das amostras atrasadas. Seja $\tilde{s}_n$ uma aproximação de s_n, onde:

$$\tilde{s}_n = -\sum_{k=1}^{p} a_k s_{n-k} \tag{3.3}$$

Então o erro total entre o valor s_n e o valor predito $\tilde{s}_n$ é obtido por:

$$e_n = s_n - \tilde{s}_n = s_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{3.4}$$

onde e_n é conhecido como resíduo.

No método dos mínimos quadrados, os parâmetros a_k podem ser obtidos da minimização do erro quadrático médio ou total relativo a cada um dos parâmetros.

O erro quadrático total *E*, é calculado pela Equação (3.5)

$$E = \sum_n e_n^2 = \sum_n \left(s_n + \sum_{k=1}^{p} \left(a_k s_{n-k}\right)\right)^2 \tag{3.5}$$

A faixa do somatório na Equação (3.5) não é especificada temporariamente. Primeiro minimizaremos o erro sem esta especificação. Ele é minimizado fazendo-se:

$$\frac{\partial E}{\partial a_i} = 0, \qquad 1 \le i \le p$$

derivando a Equação (3.5) em relação a a_i, tem-se:

$$\frac{\partial E}{\partial a_i} = 2\sum_n \left\{ \left[s_n + \sum_{k=1}^{p} a_k s_{n-k} \right] . \left[0 + \frac{\partial}{\partial a_i} \left(\sum_{k=1}^{p} a_k s_{n-k} \right) \right] \right\}$$

$$= 2\sum_n \left\{ \left[s_n + \sum_{k=1}^{p} a_k s_{n-k} \right] \left[a_k s_{n-i} \right] \right\} = 0$$

$$= \sum_n \left[a_k s_n s_{n-i} + a_k \sum_{k=1}^{p} a_k s_{n-k} s_{n-i} \right] = 0$$

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-i} = -\sum_n s_n s_{n-i} \qquad 1 \le i \le p \tag{3.7}$$

As Equações (3.7) formam um conjunto de p equações e p incógnitas. Estas equações podem ser resolvidas para os coeficientes do preditor $\{a_k, 1 \leq k \leq p\}$ que minimizam E na Equação (3.5).

O erro quadrático total mínimo, denotado por E_p é obtido expandindo a Equação (3.5) e substituindo na Equação (3.7), resultando em:

$$E_p = \sum_n s_n^2 + \sum_{k=1}^{p} a_k \sum_n s_n s_{n-k} \tag{3.8}$$

Especificando os limites do somatório sobre n nas Equações (3.5), (3.7) e (3.8). E analisando um caso de interesse para o cálculo dos parâmetros. O método utilizado para obter esses parâmetros é o método da autocorrelação.

3.2.1 - Método da autocorrelação

Assumindo que o erro na Equação (3.5) é minimizado no intervalo $-\infty \leq n \leq \infty$. Então as Equações (3.7) e (3.8) reduzem para:

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i), \quad 1 \leq i \leq p \tag{3.9}$$

$$E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{3.10}$$

Pode-se notar na Equação (3.10) que o erro mínimo total consiste de um componente fixo e outro que depende dos coeficientes do preditor.

R(i) nas Equações (3.9) e (3.10) é a função autocorrelação do sinal s_n, é obtida por:

$$R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n-i} \tag{3.11}$$

Pode-se notar que *R(i)* é uma função par de *i*, ou seja,

$$R(i) = R(-i) \tag{3.12}$$

Desde que os coeficientes de *R(i-k)* formam uma matriz autocorrelação, este método é chamado método da autocorrelação.

O interesse está em s_n sobre um intervalo finito ou ele é conhecido apenas em um intervalo finito. Um método comum é multiplicar o sinal s_n por uma janela h_n para obter o sinal s_n' que é zero fora do intervalo $0 \leq n \leq N-1$.

$$s_n' = \begin{cases} s_n h_n & ; 0 \leq n \leq N-1 \\ 0 & ; \text{fora} \end{cases} \tag{3.13a}$$

onde h_n é a janela de Hamming.

$$h_n = 0{,}54 - 0{,}46.\cos\left(\frac{2\pi n}{N-1}\right) \; ; \quad 0 \leq n \leq N-1 \tag{3.13b}$$

A função autocorrelação é obtida por

$$R(i) \sum_{n=0}^{N-i-1} s'_n \, s'_{n+1} \quad ; \quad i \geq 0 \tag{3.14b}$$

3.2.2 - Cálculo do ganho

É razoável esperar que o ganho, *G*, possa ser determinado igualando a energia do sinal com a energia das amostras linearmente preditadas. Isto é verdadeiro quando são feitas suposições apropriadas sobre o sinal de excitação para o sistema LPC.

A Equação (3.4) pode ser reescrita como

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + e_n \tag{3.15}$$

Comparando a Equação (3.1) com a Equação (3.15) pode-se ver que o único sinal de entrada u_n que resulta no sinal s_n como saída é $G_n u_n = e_n$. Isto é, o sinal de entrada é proporcional ao erro. Para outra entrada u_n a saída do filtro *H(z)* é diferente de s_n. Contudo se insistir que, para qualquer entrada u_n a energia no sinal de saída precisa ser igual à do sinal s_n, pode-se pelo menos especificar a energia total no sinal de entrada. Desde que o filtro *H(z)* é fixo, está claro com exposto acima, que a energia do sinal de entrada Gu_n precisa ser igual à energia total no sinal erro, que é dado por E_p na Equação (3.10).

É necessário fazer algumas suposições sobre u_n , para que se possa calcular *G* a partir das qualidades conhecidas, como os a_k's e os coeficientes de autocorrelação. Existem dois casos de interesse: um para sons sonoros e outro para sons surdos.

a) Para sons sonoros

Neste caso é razoável assumir $u_n = \delta_n$, isto é, a excitação é um impulso ou uma amostra unitária em n = 0. A saída é então a resposta ao impulso h_n ,onde:

$$h_n = -\sum_{k=1}^{p} a_k h_{n-k} + G\delta_n \tag{3.16}$$

A autocorrelação $\hat{R}(i)$ da resposta ao impulso h_n tem uma relacionamento interessante com *R(i)* do sinal s_n. Multiplicando se a Equação (3.16) por h_{n-1} e fazendo-se o somatório sobre todo *n* tem-se:

$$\hat{R}(i) = -\sum_{k=1}^{p} a_k \hat{R}(i-k) \ , \quad 1 \leq |\, i \,| \leq \infty \tag{3.17}$$

$$\hat{R}(0) = -\sum_{k=1}^{p} a_k \hat{R}(k) + G^2 \tag{3.18}$$

Da condição que a energia total em h_n e s_n precisam ser iguais, tem-se:

$$\hat{R}(0) = R(0)$$

desde que o coeficiente da autocorrelação de ordem zero é igual a energia total do sinal. Da Equação (3.19) e comparando a Equação (3.9) com a Equação (3.17) concluí-se que:

$$\hat{R}(i) = R(i), \quad 1 \leq i \leq p \tag{3.20}$$

A Equação (3.20) indica que os primeiros *p+1* coeficientes da autocorrelação da resposta ao impulso de *H(z)* são idênticos aos correspondentes coeficientes da autocorrelação do sinal.

Das Equações (3.18), (3.20) e (3.10) o ganho é igual a:

$$G^2 = E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{3.21}$$

onde G^2 é a energia total da entrada $G\delta_n$.

b) Sons surdos

Em sons surdos é mais razoável assumir que a entrada u_n é um ruído branco com média zero e variância unitária, isto é:

$E[u_n] = 0$ para todo n e $E[u_n u_{n-i}] = \delta_i$.

Se o sistema é excitado com a entrada randômica Gu_n , então:

$$g_n = -\sum_{k=1}^{p} a_k g_{n-k} + Gu_n \tag{3.22}$$

Se $\tilde{R}(i)$ é a autocorrelação da função g_n então:

$$\tilde{R}(i) = E[g_n g_{n-i}] = -\sum_{k=1}^{p} a_k E[g_{n-k} g_{n-i}] + E[Gu_n g_{n-i}] \tag{3.23}$$

$$= -\sum_{k=1}^{p} a_k \tilde{R}(i-k), \quad i \neq 0 \tag{3.24}$$

desde que $E[u_n g_{n-i}] = 0$ para $i > 0$, porque u_n não está correlacionado com o outro sinal anterior a u_n. Para $i = 0$, tem-se:

$$\tilde{R}(0) = -\sum_{k=1}^{p} \tilde{R}_k(k) + GE[u_n g_n]$$

$$= -\sum_{k=1}^{p} a_k \hat{R}(k) + G^2 \qquad (3.25)$$

desde que $E[u_n]\ g_n = E\ [u_n\ (G_n$ + termos anteriores $n)] = G^2$. Sempre que a energia na resposta para Gu_n precisar ser igual à energia do sinal, tem-se:

$\tilde{R}(i) = R(i)$ ou

ou

$$G^2 = R(0) + \sum_{k=1}^{p} a_k R(k)$$

como sons sonoros.

3.2.3 - Cálculo dos parâmetros do preditor

Os coeficientes do preditor a_k, $1 \leq k \leq p$ podem ser calculados resolvendo um conjunto de p equações com p incógnitas. Essas equações são as Equações (3.9). Existem alguns métodos padrões para resolver essas equações, como redução de Gauss ou método da eliminação e método da redução de Crout [23]. Esses métodos gerais requerem $\frac{p^3}{3} + Op^2$ operações (multiplicações e divisões) e p^2 locais de memória.

É possível reduzir o tempo de computação e armazenamento na resolução da Equação (3.9) devido à sua forma especial. Estas equações podem ser expandidas na forma matricial como:

$$\begin{bmatrix} R_0 & R_1 & R_2 & \dots & R_{p-1} \\ R_1 & R_0 & R_1 & \dots & R_{p-2} \\ R_2 & R_1 & R_0 & \dots & R_{p-3} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ R_{p'-1} & R_{p-2} & R_{p-3} & \dots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ . \\ a_p \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ . \\ . \\ . \\ R_p \end{bmatrix} \tag{3.26}$$

Note que a matriz autocorrelação *p x p* é simétrica e os elementos ao longo de qualquer diagonal são idênticos (isto é, uma matriz Toeplitz). Levinson derivou um procedimento recursivo elegante para resolver este tipo de equação. Esse procedimento foi posteriormente reformulado. O método de Levinson [23] assume que o vetor coluna do lado direito da Equação (3.26) assume que o vetor coluna do lado direito (3.26) é um vetor coluna geral. Usando o fato que esse vetor coluna contém os mesmos elementos encontrados na matriz de autocorrelação, outro método atribuído para Durbain [23] torna o método duas vezes mais rápido que o algoritmo de Levinson. O método requer somente *2p* localizações e $p^2 + Op$ operações. O método recursivo de Durbain pode ser especificado da seguinte maneira:

$$E_0 = R(0) \tag{3.27a}$$

$$k_i = \frac{-\left[R(i) + \sum_{j=1}^{i-1} a_j^{(j-1)} R(i-j)\right]}{E_{i-1}} \tag{3.27b}$$

$$a_i^{(i)} = k_i \tag{3.27c}$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \quad , 1 \leq \text{j} \leq \text{i-1} \tag{3.27d}$$

$$E_i = (1 - k_i^2) E_{i-1} \tag{2.27e}$$

As Equações (3.27a) – (3.27e) são resolvidas recursivamente para i = 1, 2, 3, ...,p. A solução final é obtida por:

$$a_j = a_j^{(p)}, \quad 1 \leq j \leq p \tag{3.27f}$$

Veja que na obtenção da solução para um preditor de ordem p, é realmente comutada a solução para todos os preditores de ordem inferior a *p*.

É preciso enfatizar que para muitas aplicações, a solução das Equações (3.9) não constitui o maior esforço computacional. O cálculo dos coeficientes da autocorrelação requer *pN* operações, que podem ser determinantes no tempo de computação se $N >> p$, como é o caso freqüentemente.

A solução da Equação (3.26) não é afetada se todos os coeficientes de autocorrelação são divididos por uma constante. Em particular, se todos *R(i)* são normalizados pela divisão por *R(0)* encontra-se os coeficientes normalizados de autocorrelação *r(i)*:

$$r(i) = \frac{R(i)}{R(0)} \tag{3.28}$$

os quais têm propriedade que $|r(i)| \leq 1, 0 \leq i \leq p$.

Se os coeficientes da autocorrelação são normalizados pela Equação (3.28), então o erro mínimo E_i é também dividido por *R(0)*. Chama-se a quantidade resultante de erro normalizado de V_i.

$$V_i = \frac{E_i}{R(0)} = 1 + \sum_{k=1}^{i} a_k r(k) \tag{3.29}$$

com

$$0 \le V_i \le 1 \ , \ i \ge 0 \tag{3.30}$$

Das Equações (3.27e) e (3.29), o erro normalizado para $i = p$ é obtido por:

$$V_p = \prod_{i=1}^{p} (i - k_i^2) \tag{3.31}$$

onde as quantidades k_i estão na faixa $-1 \le k_i \le 1$.

A condição dos parâmetros k_i é importante desde que é uma condição necessária e suficiente para que todas as raízes do polinômio *A(z)* estejam dentro de um círculo unitário, garantindo assim a estabilidade do sistema *H(z).*

É preciso ser notado que a garantia teórica de estabilidade para o método da autocorrelação pode não ser válida na prática se a função autocorrelação é calculada sem precisão suficiente. Markel e Gray [23] mostraram que esses efeitos indesejáveis podem ser minimizados pela pré-enfatização do sinal de voz, para tornar o espectro tão plano quanto possível.

O algoritmo de Durbain [23] permite um teste conveniente para a estabilidade desde que é necessário e suficiente que os parâmetros k_i precisam satisfazer a condição:

$$-1 \le k_i \le 1 \tag{3.32}$$

Então, se no processo de determinar os coeficientes do preditor $\{a_i\}$ algumas quantidades k_i violam a Equação (3.32) é sabido que existem raízes de *A(z)* fora do círculo unitário.

3.3 - Interpretações no Domínio da Freqüência de Análises em LPC

Até o momento o método de predição linear foi derivado de formulações no domínio do tempo. É mostrado neste capítulo que algumas equações podem ser obtidas no domínio da freqüência.

3.3.1 - Formulações no domínio da freqüência

O erro e_n entre o sinal real e o sinal predito é obtido pela Equação (3.4). Aplicando-se a transformada Z nesta equação obtém-se:

$$E(z) = \left[1 + \sum_{k=1}^{p} a_k z^{-k}\right] S(z) = A(z)S(z) \tag{3.33}$$

onde *A(z)* é o filtro inverso e *E(z)* são as transformadas *Z* de e_n e s_n, respectivamente. Porém e_n pode ser visto como o resultado de passar s_n pelo filtro *A(z)*.Aplicando o teorema de Parseval, o erro total é obtido por:

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| e^{j\omega} \right|^2 d\omega \tag{3.34}$$

onde $E(e^{j\omega})$ é obtido calculando $E(z)$ no círculo unitário $z = e^{j\omega}$. Denotando o espectro de potência do sinal s_n por $P(\omega)$, onde

$$P(\omega) = \left|S(e^{j\omega})\right|^2 \tag{3.35}$$

Da Equação (3.33) ao quadrado vem

$$|E(e^{j\omega})|^2 = |A(e^{j\omega}).S(e^{j\omega})|^2 = |A(e^{j\omega})|^2 . |S(e^{j\omega})|^2 \tag{3.35a}$$

$$|E(e^{j\omega})|^2 = P(\omega).A(e^{j\omega}).A(e^{-j\omega}) \tag{3.35b}$$

Substituindo-se a Equação (3.35b) na Equação (3.34), vem:

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(\omega)A(e^{j\omega})A(e^{-j\omega})d\omega \tag{3.36}$$

Pode ser obtido o mesmo resultado no domínio do tempo se E é minimizado aplicando a Equação (3.6) na Equação (3.36) e a autocorrelação R(i) obtida do espectro P(ω) pela transformada inversa de Fourier:

$$R(i) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(\omega)\cos(i\omega)d\omega \tag{3.37}$$

Note que na Equação (3.37) a transformada cosseno é adequada desde que P(ω) é real e par. O erro quadrático mínimo pode ser obtido substituindo-se as Equações (3.9) e (3.37) na Equação (3.36) o que resulta na Equação (3.10).

3.3.2 - Interpretação no domínio da freqüência do erro médio quadrático de predição

É examinado como o espectro do sinal *P(ω)* é aproximado pelo espectro do modelo tudo-polo, que é denotado por *P(ω)*. Da Equação (3.2).

$$P(\omega) = |H(e^{j\omega})|^2 = \frac{G^2}{|A(e^{j\omega})|^2} = \tag{3.38}$$

$$= \frac{G^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-j\omega k}\right|^2}$$

Das Equações (3.33) e (3.35) tem-se:

$$P(\omega) = \frac{\left|E(e^{j\omega})^2\right|}{\left|A(e^{j\omega})\right|^2} \tag{3.39}$$

Comparando-se as Equações (3.38) e (3.39) pode-se ver que se *P(ω)* está sendo modelado por $\hat{P}(\omega)$, então o espectro de potência do erro $|E(e^{j\omega})|^2$ está sendo modelado por espectro do plano igual a G^2. Isto significa que o sinal erro está sendo aproximado por

outro sinal que tem espectro plano, tal como um impulso unitário, ruído branco, ou algum outro sinal com espectro plano.

De (3.34), (3.38) e (3.39) o erro total pode ser escrito como:

$$E = \frac{G^2}{2\pi}\int_{-\pi}^{\pi}\frac{P(\omega)}{\hat{P}(\omega)}d\omega \tag{3.40}$$

Desde que o integrando da Equação (3.40) é positivo, significa que minimizar *E* é equivalente a minimizar a integral da razão do espectro de energia do segmento de voz *P(ω)* para a magnitude ao quadrado da resposta em freqüência do sistema linear no modelo de produção de voz $\hat{P}(\omega)$.

A Equação (3.20) mostra que a função autocorrelação do segmento de voz s_n e a função autocorrelação $\hat{R}(i)$ da resposta ao impulso h_n (correspondente a *H(z)*) são iguais para os primeiros *p+1* valores. Desde que *P(ω)* e $\hat{P}(\omega)$ são transformadas de Fourier de *R(i)* e $\hat{R}(i)$, respectivamente, segue que incrementando-se a ordem do modelo p, aumenta-se a faixa em que *R(i)* e $\hat{R}(i)$ são iguais, resultando em uma melhor adaptação de $\hat{P}(\omega)$ para *P(ω)*. No limite, quando $p \to \infty$, $\hat{R}(i)$ torna-se idêntico a *R(i)* para todo *i*, e portanto os dois espectros tornam-se idênticos:

$$\hat{P}(\omega) = P(\omega) \quad \text{quando } p \to \infty \tag{3.41}$$

A Equação (3.41) mostra que se p é suficientemente grande pode-se aproximar o espectro do sinal, com um erro arbitrariamente pequeno, com o modelo tudo-polo *H(z)*.

É interessante notar que embora a Equação (3.41) mostre que quando $p \to \infty$ tem-se $|H(e^{j\omega})|^2 = |S(e^{j\omega})|^2$, não é necessariamente verdadeiro que $H(e^{j\omega}) = S(e^{j\omega})$, isto é, a resposta em freqüência do modelo não necessita ser igual à transformada de Fourier do sinal. Isto porque $S(e^{j\omega})$ não necessita ter fase mínima, ao passo que $H(e^{j\omega})$ requer fase mínima, desde que ele é a função de transferência de um filtro tudo-polo com os pólos dentro do círculo unitário.

3.4 - Medidas de Distorção

Considere o sistema mostrado na Figura 3.5 onde tem-se vetor de entrada *x* e sua reprodução *y*. Pode-se definir a distorção *(x,y)* como a designação de um número positivo para a diferença obtida quando o vetor de entrada é substituído pelo vetor reprodução *y*.

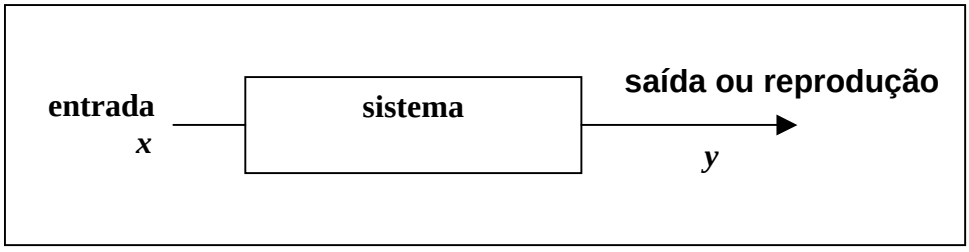


Figura 3.2 – Sistema com uma entrada *x* e uma reprodução *y*.

Uma medida de distorção para ser útil precisa ter duas características. Primeiro, ela precisa ser tratável no sentido em que possa produzir análises matemáticas acessíveis. Segundo, precisa ser subjetivamente significante de modo que as diferenças nos valores da distorção indiquem diferenças similares na qualidade da voz.

Neste trabalho evita-se usar o termo “distância”, pois falando estritamente, a “distância” requer as seguintes características [20]:

1) não negatividade: $d(x,y) \geq 0$, com igualdade somente para $x = y$;

2) simetría: $d(x,y) = d(y,x)$, *e;*

3) desigualdade triangular $d(x,z) \leq d(x,y) + d(y,z)$.

Se $d(x,y)$ é uma distorção então $d(x,y) > 0$ e se $x = y$ então $d(x,y) = 0$. Isto é, requer a primeira característica da distância.

Relacionam-se abaixo algumas das medidas de distorção mais usadas.

3.4.1 - Erro médio quadrático (mse) [19]

A medida de distorção mais comum é o erro médio quadrático

$$d_2(x,y) = \frac{1}{N}(x-y)^t(x-y) = \frac{1}{N}\sum_{k=1}^{N}(x_k - y_k)^2 \qquad (3.42)$$

onde x e y são vetores de dimensão N. A distorção é definida pela dimensão. Esta é a medida de distorção mais popular, devida principalmente a sua simplicidade e ao tratamento matemático simples.

Uma medida de distorção mais geral é definida por:

$$d_r(x,y) = \frac{1}{N}\sum_{k=1}^{n} | x_k - y_k |^r \qquad (3.43)$$

Note que a Equação (3.43) é igual a Equação (3.42) para $r = 2$. Os dois valores mais comuns de r são 1 e ∞. d_1 representa o erro médio absoluto e d_∞ tende para o erro máximo. De fato, pode ser mostrado que:

$$\lim_{r\to\infty}[d_r(x,y)]^{\frac{1}{r}} = máx\{|x_k - y_k|, 1 \le k \le N\} \tag{3.44}$$

Para a codificação de voz, d_2 tem sido a medida de distorção mais popular. Com d_1 e d_∞ sendo usadas ocasionalmente.

3.4.2 - Erro médio quadrático com peso [19]

No *mse* (d_2) assumimos que as contribuições da distorção por cada x_k são igualmente consideradas. Em geral, pesos diferentes podem ser introduzidos de modo a proporcionar que certas contribuições para a distorção sejam mais importantes que outras. Um *mse* com peso geral pode ser definida por:

$$d_w(x,y) = (x-y)^t W (x-y) \tag{3.45}$$

onde W é uma matriz de peso positiva. Se $W=N^{-1}I$, onde I é a matriz identidade, resulta que $d_w = d_2$. Uma escolha para W que é popular em muitas aplicações de modelos de classificação é $W = \Gamma^{-1}$, onde Γ é a matriz convergência do vetor randômico x.

$$\Gamma = E[(x-\bar{x})(x-\bar{x})^t] \ , \bar{x} = \varepsilon(x) \tag{3.46}$$

Neste caso d_w reduz para

$$d_w(x,y) = (x-y)^t \Gamma^{-1} (x-y) \tag{3.47}$$

e é conhecida como distância de Mahalanobis [4].

Se a matriz *W*, além de positiva, for simétrica (como no caso da distância de Mahalanobis) a matriz *W* pode ser escrita como:

$$W = P^t . P \tag{3.48}$$

Os vetores *x* e *y* podem ser transformados em um novo conjunto de vetores x' e y', onde

$$x' = P_x \text{ e } y' = P_y \tag{3.49a}$$

$$d_w(x,y) = (x-y)^t P^t P (x-y) = (P_x - P_y)^t (P_x - P_y)$$

$$= (x`-y`)^t (x`-y`) = d_2(x`, y`) \tag{3.49b}$$

A Equação (3.49b) que o *mse* com peso entre os valores originais é igual a *mse* dos valores transformados.

3.4.3 - Medidas de distorção usando LPC [14, 18, 19 ,23]

No reconhecimento do locutor é necessário comparar de maneira quantitativa e computacionalmente eficiente, dois quadros de voz que a análise de LPC cede conjuntos de

coeficientes LPC diferentes. Então procurou-se um medida *d(x,y)* onde *d* é a distorção entre os quadros de voz com os conjuntos de parâmetros LPC *x={1, x(1), x(2), ..., x(p)}* e *y={1, y(1), y(2), ..., y(p)}*.

Uma medida de distorção sofisticada foi proposta por Itakura [4]. Essa medida pode ser obtida pelo seguinte entendimento. Pode ser questionado que devido ao ruído, bem como a inexatidão do modelo de predição linear da voz, não é possível medir os verdadeiros coeficientes LPC associados com um segmento de voz. É possível somente estimar os LPC's básicos para o segmento de voz. Assumindo que se tem um segmento de voz com coeficientes LPC's estimados *x*. O problema é determinar a probabilidade que x é de um segmento de voz com coeficientes LPC's verdadeiros *y*. Uma vez que esta probabilidade é determinada, pode ser obtida uma medida efetiva para avaliar a dissimilaridade.

Foi mostrado por Rabiner [23] que a distribuição de probabilidade que governa as estimações de *y* é uma distribuição multidimensional Gaussiana com média *y* e matriz covariância λ definida como

$$\lambda = \frac{1}{N} R^{-1}(xRx^t) \tag{3.50}$$

onde *R* é a (*p+1* x *p+1*) matriz autocorrelação de voz. *N* é o comprimento do quadro em amostras, e *t* indica a transposta. Então a probabilidade de obter a estimação *x* quando os coeficientes LPC básicos são *y* é:

$$P(x/y) = (2\pi)^{\frac{-p}{2}} \mid \lambda \mid^{\frac{1}{2}} \exp[-0{,}5(x-y)\lambda^{-1}(x-y)^t] \tag{3.51}$$

onde $|\lambda|$ é o determinante da matriz λ.

Uma medida de distância apropriada é obtida calculando-se o logaritmo da Equação (3.51). A medida de distorção resultante é:

$$d(x,y)=(x-y)\lambda^{-1}(x-y)^t \tag{3.52}$$

Multiplicando-se os dois lados da Equação (3.50) por λ^{-1} obtendo-se:

$$I = \frac{R^{-1}}{N}(xRx^t)\lambda^{-1} \tag{3.53}$$

onde I é a matriz identidade. Multiplicando-se os dois lados da Equação (3.53) por R tem-se:

$$\lambda^{-1} = \frac{NR}{xRx^t} \tag{3.54}$$

Substituindo-se a Equação (3.54) na Equação (3.52) obtém-se:

$$d(x,y) = (x-y)\frac{NR}{xRx^{-1}}(x-y)^t \tag{3.55}$$

É visto facilmente que quanto maior a probabilidade que x tenha vindo da distribuição dos coeficientes LPC's y, menor é a distância calculada usando-se a Equação (3.55). Devido a considerações computacionais Itakura propôs a seguinte medida de distorção:

$$d`(x,y) = \log\left[\frac{yRy^t}{xRx^t}\right] \tag{3.56}$$

Uma medida de distorção alternativa foi proposta por Itakura e Saíto [4], ela deriva do princípio da máxima probabilidade. A forma modificada dessa distorção entre um vetor de coeficientes do preditor *x* e outro vetor de coeficientes *y* obtida por:

$$d(x,y) = (x-y)R_x(x-y) \tag{3.57}$$

$$\text{onde} \quad R_x = \left\{\frac{r(i-k)}{r(0)}; 0 \le i,k \le p-1\right\} \tag{3.58}$$

é a matriz autocorrelação normalizada cujos os coeficientes *r(i-k)* foram usados no cálculo dos coeficientes do preditor *x* na Equação (3.27). Desde que os coeficientes autocorrelação nessa Equação são normalizados por *r(0),* pode ser mostrado que R_x e o vetor *x* determinam unicamente um ao outro.

Uma outra derivação da medida de distorção é obtida pela equação [27]:

$$d(x,y) = \frac{yR_xy^t}{xR_xx^t} - 1 \tag{3.59}$$

Sendo R uma matriz Toeplitz o termo xRx^t pode ser escrito na forma computacionalmente eficiente:

$$xRx^t = \sum_{i=0}^{p}\sum_{j=0}^{p} x_i x_j r(i-j) \tag{3.60a}$$

$$= b(0)r(0) + 2\sum_{i=0}^{p} b(i)r(i) \qquad (3.60b)$$

onde

$$b(i) = \sum_{j=0}^{p-i} x_j x_{j+1} \qquad ,0 \leq i \leq p \qquad (3.60c)$$

3.5 – Quantização Vetorial

3.5.1 – Formulação do problema

Seja $x = (x_1,\ldots, x_k)$ um vetor de dimensão K, cujos componentes $\{x_i, i=1, \ldots,K\}$ são estimações reais de variáveis aleatórias de amplitudes contínuas. Na quantização de vetores o vetor x é quantizado como y, onde y é o valor quantizado de x. Assim a operação é definida como:

$$x = q(y), \qquad (2.61)$$

onde $q(.)$ é o operador de quantização, x é chamado de vetor de reconstrução ou vetor de saídas correspondentes a y.

Tipicamente x leva a forma de um conjunto finito de valores $\hat{A} = \{x_i , i = 1, \ldots, N\}$, onde $x_i = \{x_{i1}, x_{i2, \ldots,} x_{iK} \}$. O conjunto $\hat{A}$ é referido como "*codebook*", e $\{x_i\}$ são vetores códigos. A dimensão N do quantizador é também chamada de números de níveis, como na quantização escalar.

Para projetar um quantizador, faz-se a partição do espaço S dimensional do vetor aleatório x em N regiões ou células $R = \{R_i; i = 1, \ldots,N\}$ e associa-se a cada célula C_i

um vetor x_{ci}, seguindo um determinado critério. Tais vetores são chamados de centróides da célula. O quantizador atribui o vetor código (ou centróides) x_{ci} a y se este está em C_i, ou

$$q(y) = x_{ci} \qquad \text{se } y \in C_i \tag{3.62}$$

O processo de projetar um quantizador é também conhecido como treinamento. Um algoritmo para projetar um quantizador será descrito mais adiante.

As Figuras 3.3 e 3.4 ilustram os processos de quantização. A Figura 3.3 mostra o exemplo de uma partição bidimensional ($K = 2$) para o propósito de quantização de vetores. A região fechada pelas linhas escuras é a célula C_i. Algum vetor de entrada y que se encontra na célula C_i é quantizado como x_{ci}. As posições dos vetores códigos correspondentes para outras células são amostradas por pontos. O número total de vetores códigos no exemplo dessa Figura 3.3 é $N = 19$.

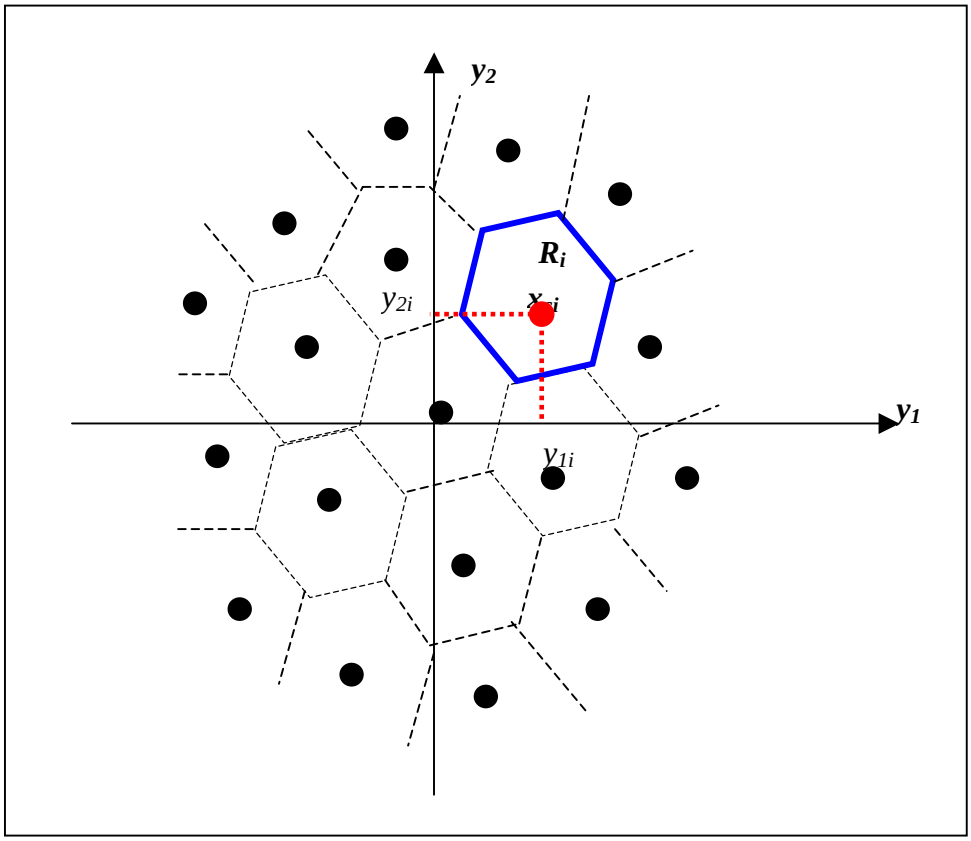


Figura 3.3 – Partição bidimensional (K=2).[9]

Para *k=1* a quantização de vetores reduz-se à quantização escalar. A Figura 3.4 mostra um exemplo da reta real para a quantização escalar. Os valores códigos (saída ou níveis de reconstrução) são mostrados por pontos. Aqui, também, algum valor de entrada *y* que se encontra-se no intervalo R_i é quantizado como x_i. O número de níveis nessa figura é *N=10*. A quantização escalar tem a propriedade de que enquanto as células podem ter dimensões diferentes, todas elas têm a mesma forma, pois são intervalos na linha real. A liberdade de ser células de várias formas, no espaço multidimensional, fornece a quantização de vetores uma vantagem sobre a quantização vetorial escalar.

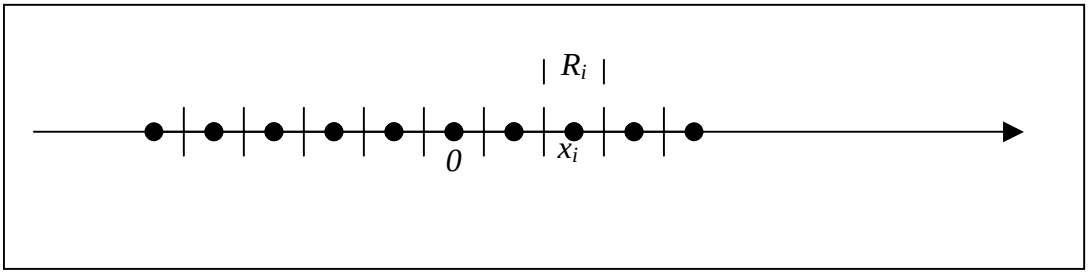


FIGURA 3.4 - Partição unidimencional (*k=1*) [9].

3.5.2 – Quantização ótima

Um quantizador de *N* níveis é ótimo se ele minimiza a distorção esperada, isto é, q^* é ótimo se para todos os outros quantizadores *q*, tendo *N* vetores de reprodução, $D(q^*) \leq D(q)$. Um quantizador é localmente ótimo se $D(q)$ é somente o mínimo local, isto é, pequenas mudanças em *q* causam um aumento na distorção. A meta de projeto de uma quantizador vetorial é obter uma quantizador ótimo se for possível, porém se não for, deve-se obter o menor o ótimo local [9]. Em alguns casos, ótimos globais podem ser obtidos analiticamente, ou então, por algum método exaustivo sobre os ótimos locais.

Na quantização particiona-se o espaço em N células que são associados os respectivos centróides. O quantizador então apontar para o centróide x_{ci} se x (entrada do sistema) está em C_i. O quantizador é ótimo se a distorção é minimizada nos N níveis. Existem duas condições necessárias para a otimização [6]:

1º.) A Quantização deve ser realizada usando uma regra de distorção mínima ou de seleção dos vizinhos mais próximos:

$$q(x) = y_i \Leftrightarrow d(x, y_i) \leq d(x, y_j), \tag{3.62}$$

onde $j \neq i$, $1 \leq j \leq N$;

2º.) Cada x_{ci} deve ser escolhido de modo a minimizar a distorção média dentro da célula R_i. Na prática são dados os vetores de treinamento x. Um subconjunto com M_i desses vetores pertence a célula R_i . A distorção média é obtida por:

$$D_i = \frac{1}{M_i} \sum_{x \in C_1} d(x, x_{ci}) \tag{3.63}$$

Para a distorção média quadrática, com ou menos pesos, é possível verificar que D_i é minimizado quando x_{ci} é simplesmente a média aritmética dos vetores contidos em R_i, ou seja:

$$x_{ci} = \frac{1}{M_i} \sum_{x \in C_i} x \tag{3.64}$$

Para a distorção de Itakura–Saito, y_i é calculado inicialmente calculando-se as médias das autocorrelações:

$$R_{y_i} = \frac{1}{M_i} \sum_{x \in C_i} R_x(k), \quad k = 0,..., p \tag{3.65}$$

com $R_x(k)$ normalizados, e a seguir obtém-se y_i como a solução da Equação (3.9), sendo R_{y_i} os coeficientes de autocorrelação.

Vários algoritmos tem sido propostos dentre eles o Algoritmo das k-Médias [6], Algoritmos baseados nos Métodos de Lloyd e Algoritmo LBG. Este último é explicado a seguir.

3.5.3 – Algoritmo LBG

O algoritmo LBG foi utilizado neste trabalho para projetar o quantizador vetorial. O algoritmo LBG, assim denominado pelas iniciais dos nomes de seus criadores, Llinde, Buzo e Gray [9], é um modelo bastante apropriado para a utilização nos processos que envolvem o sinal da fala. O objetivo geral deste algoritmo é achar um conjunto de N vetores, chamados de alfabeto de reprodução, dentro dos quais todos os vetores característicos da seqüência de treinamento possam ser quantizados com distorção mínima [4]. A seqüência de treinamento consiste de uma seqüência de vetores resultantes, normalmente, de um filtro inverso normalizado. Tais saídas do filtro correspondem aos coeficientes preditos linearmente, ou coeficientes LPC.

Na seção 4.4.2 do capítulo 4 deste trabalho é mostrado o algoritmo LBG.

3.6 – Considerações Finais deste Capítulo

Foi apresentado neste capítulo uma breve visão do processamento digital de sinais da fala e as técnicas de processamento digital de sinais da fala e as técnicas de

processamento digital de sinais, tais como LPC, medida de Distorção e quantização vetorial.

E ainda que os vetores característicos podem ser encontrados resolvendo as Equações (3.12) para calcular a autocorrelação e usando o método de Durbain (Equações (3.27)) para encontrar os coeficientes do preditor.

O modelo tudo-polo *H(z)* aproxima-se do espectro do sinal quando *p* aumenta.

CAPÍTULO IV

O CLASSIFICADOR POLINOMIAL NO RECONHECIMENTO DO LOCUTOR

A técnica polinomial para classificação tem sido usada por *K. Fukunaga* durante vários. Recentemente alguns autores como *W.M. Campbell,* têm trabalhado com o classificador polinomial para reconhecimento do locutor pela sua voz. O método está ainda pouco explorado principalmente quanto aos tipos e ordem de polinômios. Neste capítulo é discutido a estrutura do classificador, a sua eficiência na separabilidade e na classificação, objetivando aplicações em reconhecimento do locutor.

4.1 - Introdução

Freqüentemente depara-se com o problema de classificar os elementos pertencentes a uma ou mais classes de acordo com suas características. Este tipo de problema é encontrado no dia a dia. O ser humano tem a capacidade de fazer tal tarefa com muita facilidade. A separação de frutas como maçã ou laranja é facilmente reconhecida

pelo ser humano, pois as mesmas apresentam formas e cores diferentes. Assim é com a maioria dos objetos encontrados na natureza, sempre se faz a classificação de objetos em classes de acordo com a similaridade. Os métodos de classificação podem ser classificados em duas classes:

a- Supervisionado

Os métodos supervisionados indicam ao processo o número de classes e a quais classes os elementos pertencem. Estes são os principais tipos de classificadores usados e dentre eles pode se citar: os classificadores do tipo linear que usam superfície de hiperplanos, uma rede neural quando treinada com o algoritmo de *back-propagation*, os classificadores polinomiais, etc. Se, por exemplo, for necessário classificar maçãs e laranjas, é necessário fornecer a característica da maçã e da laranja.

b- Não Supervisionado

Nos classificadores não supervisionados indica-se o número de classes, mas não fornece qual classe os elementos pertencem. Os algoritmos classificam os elementos de acordo com suas similaridades. Se por exemplo, tem-se um conjunto de maças e laranjas, o classificador faz a separação dos elementos de acordo com suas similaridades. Isto é, o classificador tem um aprendizado sobre o sistema ao mesmo tempo em que classifica os elementos, é um auto-aprendizado. Não é difícil aprender a diferença entre uma laranja e uma maçã. Os principais classificadores não supervisionados são: a quantização vetorial, o algoritmo de Kohonen em redes neurais, etc.

4.2 - Quantização

4.2.1 - Quantização escalar

Como foi visto, na seção 3.5.1 do capítulo III deste trabalho, a quantização escalar é um processo de aproximação de uma grandeza escalar, isto é de uma dimensão, como por exemplo, uma voltagem de um circuito e a temperatura de uma sala. Nesse processo que é usado em conversores A/D, o sinal *x* é aproximado pelo nível discreto mais próximo, como mostra a Figura 4.1. Os valores discretos neste exemplo são: *-3, -1, 1, 3,* que são representados pelos níveis binários *00, 01, 10* e *11*, respectivamente. Valores de *x = 0,6; 0,8 ; 1,2; 1,3* e etc. são aproximados para o nível 1.

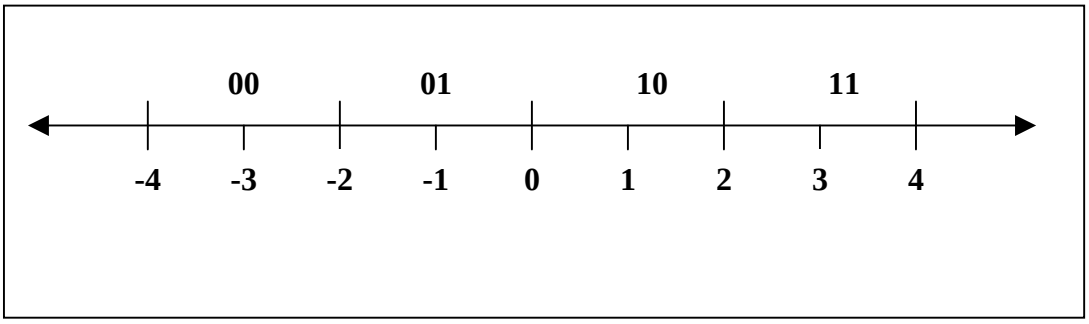


Figura 4.1 – Processo de quantização escalar no conversor A/D.

Os valores do intervalo *[-2 0)* são aproximados pelo valor *–1* ou pelo nível binário 01, valores no intervalo *[0 2)* são representados pelo valor 1 ou pelo nível binário 10 e etc.. No sistema tem-se um conjunto de números na reta real que pode ser representado por quatro níveis. O erro na aproximação é chamado erro de quantização e é obtido por,

$$e = |x - y_i| \tag{4.1}$$

Onde x é o valor que se deseja aproximar, e y_i é o nível mais próximo de x. Pode-se observar que o erro máximo no exemplo é menor ou igual a 0,5; ou seja, $e \leq 0{,}5$.

4.2.2 – Quantização vetorial

A técnica de quantização é também um método de aproximação, mas para dados apresentados na forma vetorial, como foi visto na seção 3.5.1, do capítulo III deste trabalho. Como por exemplo, se tivermos um sistema bidimensional como apresentado na Figura 3.3 da seção 3.5.1, pode-se aproximar um vetor x qualquer pelo vetor mais próximo, que representa uma classe , também chamado de centróide. Nessa figura tem-se 19 regiões e cada região é representada pelo seu centróide. Sendo assim se um vetor cair em uma região R_i, ele será aproximado pelo centróide x_{ci}.

Pelo exposto acima é possível representar os 16 centróides com 4 bits. Em um processo de armazenamento ou transmissão pode-se armazenar o conjunto de centróides, também chamado '*codebook*', no receptor e transmitir apenas o binário que representa o centróide que o valor do vetor x foi aproximado.

Mas para construir um '*codebook*', a questão se resume em:

Dado uma fonte vetorial com suas propriedades estatísticas conhecidas, uma medida de distorção e um número de vetores código (*codevectors*), encontrar um '*codebook*' e uma partição que resulta na menor distorção.

Para o caso do sinal de voz, cada quadro é convertido em um vetor e cada vetor vai compor a seqüência de treinamento,

$$S = \{ x_1, x_2, ..., x_M \}$$

O algoritmo mais usado em sinais da fala é o algoritmo LBG [17], que foi desenvolvido por Lindle, Buzo e Gray, como foi visto no capítulo III deste trabalho.Este algoritmo tem como entrada de dados: o conjunto *S*, o número de vetores código do 'codebook', e a aproximação necessária no processo. A saída são os vetores 'codebook'.

Os passos do algoritmo são:

a- Dado os vetores de S, calculando-se o vetor médio,

$$c_1 = \frac{1}{M}\sum_{m=1}^{M} X_m \qquad (4.2)$$

b- Calculando-se dois vetores a partir do vetor médio, considerando-se uma perturbação no mesmo,

$$N=1$$

$$C_i = (1+\varepsilon)c_i$$

$$C_{N+1} = (1-\varepsilon)c_i$$

c- Classifica-se todos os vetores de S, em uma das classes dos novos vetores C_i e C_{N+1} de acordo com a medida de distorção previamente definida.

d- Com os vetores classificados nas duas classes calcula-se os dois novos centróides.

e- Calcula-se quatro vetores a partir dos dois centróides, considerando-se uma perturbação no mesmo.

f- Classifica-se todos os vetores dos dois conjuntos *S*, em uma das novas quatro classes formadas,

g- Calculam os quatros novos centróides

h- O processo continua até o número de centróides desejados.

4.3 - Discriminador Linear

4.3.1 - O classificador hiperplano

Um hiperplano é definido como uma equação linear

$$v_1x_1 + v_2x_2 + v_3x_3 + \ldots + v_nx_n = p \tag{4.3}$$

Sendo que nenhum dos coeficientes $v_1, \ldots, v_n$ são nulos. A Equação (4.3) pode ser escrita pelo produto escalar,

$$x.\, v = p\,, \tag{4.4}$$

onde o vetor ,

$v = (\, v_1, v_2 \ldots\ldots v_n)^T$ é normal ao hiperplano.

Qualquer hiperplano *H* define dois semi-espaços. Um deles é definido por:

$$x.v \geq p$$

O outro semiplano é definido pela desigualdade,

$$x.v \leq p$$

e para passar do interior de um semiplano para o interior do outro tem-se que cruzar obrigatoriamente o hiperplano,

$$x.v = p$$

Observe que o hiperplano, $3x + 4y = 2$, separa os pontos $(0,0)^T$, e $(1,1)^T$, pois 3.0 + 4.0 < 2 e $3.1 + 4.1 > 2$.

4.3.2 - Função discriminante linear para duas dimensões

Um outro tipo de classificador é usar uma função discriminante linear definida por:

$$g(x) = w^t.x + w_o \qquad (4.5)$$

onde w é chamado de *vetor de pesos* e w_o é chamado de *limiar de pesos*

A função discriminante $g(x)$ classifica os elementos em duas regiões ou classe c_1 e c_2. Se $g(x) > 0$, x pertence à classe c_1 e se $g(x) < 0$, x pertence à classe c_2.

A função da Equação (4.5) é similar a Equação (4.3) e o hiperplano que define o limiar das duas classes ou regiões pode ser determinado pegando um ponto x_1 e x_2 ambos pertencentes à superfície de decisão, então:

$$w^t x_1 + w_o = w^t x_2 + w_o$$

$$\text{ou} \qquad w^t(x_1 - x_2) = 0$$

pode-se perceber que *w* é normal a qualquer vetor pertencente ao hiperplano.

4.3.3 - Função discriminante linear para mais de duas dimensões

Para o caso de ter mais de duas classes ou regiões, o processo de separação utilizando hiperplanos pode ser usado, mas leva a regiões não definidas, como mostra a Figura 4.2.

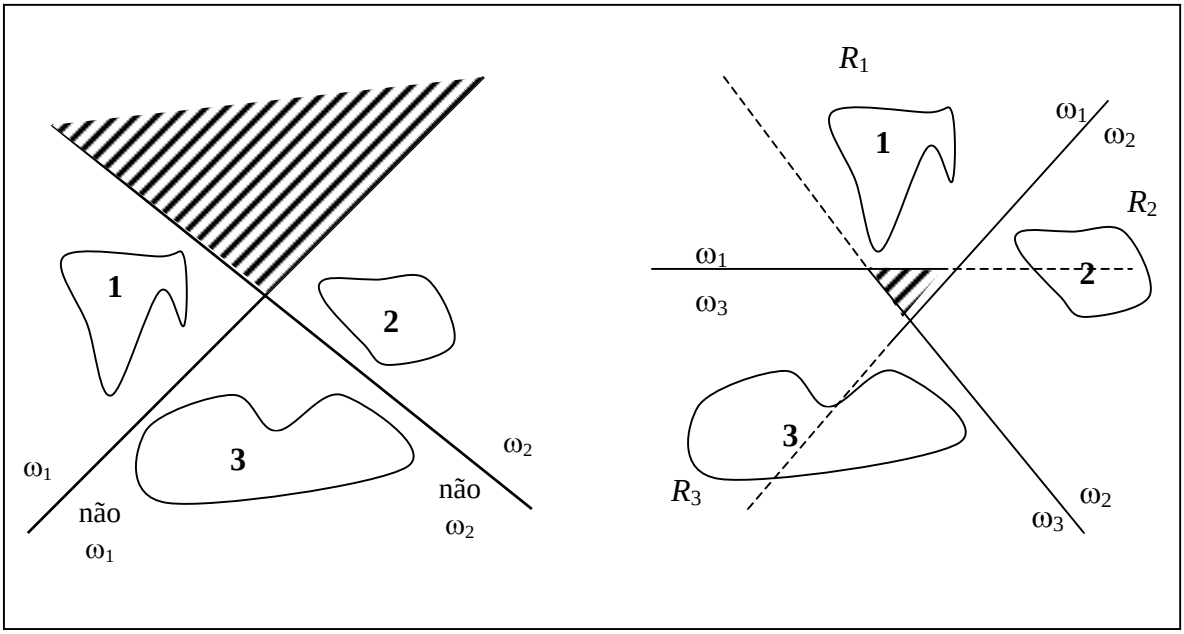


Figura 4.2 - Regiões lineares com três classes.

Para resolver este problema um processo muito usado é classificar o vetor *x* à classe mais próxima de acordo com a distância. Outro processo é usar a função discriminante para cada classe, como:

$$g_i(x) = w^t.x + w_{io} \qquad i = 1,..., c.$$

O vetor x será classificado na classe i se

$$g_i(x) > g_j(x) \qquad \text{para todo } j \neq i.$$

Se tivermos c classes teremos c funções discriminantes.

4.4 – O Classificador Polinomial

4.4.1 – Função discriminante

O classificador linear poderá ser generalizado, construindo se a função discriminante for quadrática, então:

$$f(x) = w_o + \sum_{i=1}^{n} w_i x_i + \sum_{i=}^{n} \sum_{j=1}^{n} w_{ij} x_i x_j \qquad (4.6)$$

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 generalizada, tal que para um classificador polinomial de terceira ordem,

$$f(x) = w_o + \sum_{i=1}^{n} w_i x_i + \sum_{i=}^{n} \sum_{j=1}^{n} \sum_{k=0}^{n} w_{ijk} x_i x_j x_k \qquad (4.7)$$

ou

$$f(w,x) = \sum_{l=0}^{m} w_i g_i(x) \tag{4.8}$$

onde $g_i(x)$ é uma função polinomial previamente definida e w_i são coeficientes a serem determinados. A Equação (4.8) pode ser escrita na forma matricial como,

$$f(w,x) = w^t p(x) \tag{4.9}$$

onde $p(x)$ é o polinômio cuja descrição é fornecida abaixo.

4.4.2 - A estrutura do classificador polinomial

O classificador polinomial se resume inicialmente em gerar uma matriz de pesos w que represente cada locutor. O sinal de voz do locutor a ser reconhecido é como de praxe dividido em quadros, onde são extraídos os vetores característicos de cada quadro e em seguida calcula-se um valor médio s_i para cada locutor,

$$s_i = \frac{1}{N} \sum_{j=1}^{N} f(w_i, x_j) \tag{4.10}$$

onde a função:

$$f(w,x) = w^t p(x)$$

e $p(x)$ é o vetor polinômio de grau K.

a- Como exemplo para $K=2$, e o ponto $x = (x_1,x_2)$

$$p(x) = [\ 1\ x_1\ x_2\ x_1^2\ x_1x_2\ x_2^2]$$

b- Para $K= 3$ e $x = (x_1,x_2)$, o polinômio é dado por,

$$p(x) = [\ 1\ x_1\ x_2\ x_1^2\ x_1x_2\ x_2^2\ x_1^3\ x_1^2x_2\ \ x_2^2x_1\ \ x_2^3]$$

c- Para K = 3 e $x = (\ x_1\ x_2\ ...\ x_{12})$ têm-se 455 monômios,

Têm-se 13 monômios de primeira ordem que são,

$1\ x_1\ ...\ x_{12}$

Têm-se 78 monômios de segunda ordem que são,

$x_1^2 \quad x_1x_2 \quad .... \ x_1x_{12}$

$x_2x_1 \quad x_2^2 \ ... \ x_2x_{12}$

.

.

.

$x_{12}x_1 \ \ x_{12}x_2 \ \ ... \ \ x_{12}^2$

Têm-se 114 termos de terceira ordem do tipo,

$x_1^3 \quad x_1^2x_2 \quad x_1^2x_3 \ \ ... \ \ x_1^2x_{12}$

$x_2^2x_1 \quad x_2^3 \quad x_2^2x_3 \ \ ... \ \ x_2^2x_{12}$

.

.

.

$x_{12}^2 x_1$ ... x_{12}^3

Têm-se 220 combinações dos 12 elementos do vetor $x = (x_1\ x_2\ ...\ x_{12})$ tomados 3 a 3, do tipo,

$x_1x_2x_3$ $x_1x_2x_3$ ...

$x_2x_3x_4$ $x_2x_3x_5$ ...

.

.

.

etc.

O total de termos do polinômio é 13 + 78 + 144 + 220 = 455.

Assim pode-se encontrar uma fórmula em função da dimensão do vetor, *n*, e número de elementos do polinômio, *g*, utilizando-se técnicas de análise combinatória.

Para $K = 2$:

$$\dim(p(x)) = \frac{n^2 + 3n + 2}{2} \tag{4.11}$$

Para $K = 3$:

$$\dim(p(x)) = \frac{n^3 + 6n^2 + 11n + 6}{6} \tag{4.12}$$

4.5 – Considerações Finais deste Capítulo

Neste capítulo foi visto as propriedades da separabilidade e da classificação dos quantizadores vetoriais e classificadores polinomiais. Foi verificada também como são construídas as funções discriminantes polinomiais para a segunda e a terceira ordem.

Foram apresentados também os tipos de expansões polinomiais. É importante ressaltar que não foi explorado neste trabalho as diferentes expansões associadas a sua eficiência na classificação. Porém é sabido que os polinomiais de grau superiores são mais eficientes.

Fica para o próximo capítulo, que é tratado certas situações, como a determinação dos vetores classificadores.

CAPÍTULO V

RECONHECIMENTO E VERIFICAÇÃO DO LOCUTOR USANDO CLASSIFICADOR POLINOMIAL E QUANTIZAÇÃO VETORIAL

5.1 – Introdução

No capítulo IV foram analisadas as técnicas de quantização vetorial e a classificação polinomial. Neste capítulo é usado o classificador polinomial para reconhecer o locutor a partir do vetor de pesos obtidos pelos "*codebooks*". Para cada locutor do banco de dados é gravado um texto, e constrói-se o seu "*codebooks*". A partir daí determina-se um "*speaker model"*, ou modelo do locutor e, que funcionam como um fator peso para cada uma dos locutores parâmetros. E posteriormente calcula-se o "*score*", que classifica os locutores e realiza-se as comparações necessárias para obtém-se a verificação e o reconhecimento.

5.2 - Banco de Dados dos Sinais de Cada Locutor

O banco de dados desenvolvido para os testes de reconhecimento e verificação de locutor, foi obtido apartir de 3 amostras de 30 pessoas sendo 14 do sexo masculino e 16 do sexo feminino, cuja a faixa etária é 18 à 40 anos.

Todas as gravações foram feitas em sala fechada, com ruído ambiente normal, usando um intervalo de 5 minutos em média entre uma gravação e outra, para os arquivos modelos (ou parâmetros) e um intervalo de15 minutos para a amostra de análise.

Foram utilizadas frases numéricas de conteúdos idênticos para todas amostras em cada locutor, isto é, cada locutor pronunciou três vezes a contagem natural de 0 (zero) à 9 (nove). O sinal sonoro foi capturado por um microfone de *eletreto*, tipo *hadset*, amostrado e digitalizado por uma placa compatível com *Sound Blaster* a uma taxa de amostragem de 11025 Hz, com 16bits por amostra e em Mono.

O tempo de gravação, o tamanho do arquivo e o sexo do locutor estão descritos na tabela 5.1. Nesta tabela foi adotado a seguinte nomenclatura, *LX_Y.wav* é o arquivo em formato *wave*, obtido na Y-ésima gravação do Locutor X, por exemplo, o arquivo *L5_3.wav* foi captado na terceira pronuncia do locutor 5.

Tabela 5.1- Descrição do banco de dados.

Nome do Arquivo	**Sexo**	**Tempo de Gravação (s)**	**Tamanho (Kb)**	**Nome do *Arquivo***	**Sexo**	**Tempo de Gravação (s)**	**Tamanho (Kb)**
L1_1.wav	M	9,965	214	***L16_1.wav***	M	6,208	133
L1_2.wav		11,547	248	***L16_2.wav***		6,614	142
L1_3.wav		5,902	127	***L16_3.wav***		6,541	140
L2_1.wav	F	12,089	260	***L17_1.wav***	F	6,564	141
L2_2.wav		12,462	268	***L17_2.wav***		5,591	120
L2_3.wav		12,471	268	***L17_3.wav***		5,981	128
L3_1.wav	M	9,854	212	***L18_1.wav***	F	6,832	147
L3_2.wav		6,787	146	***L18_2.wav***		6,218	133
L3_3.wav		6,687	144	***L18_3.wav***		6,063	130
L4_1.wav	F	15,855	341	***L19_1.wav***	F	5,890	126
L4_2.wav		14,117	304	***L19_2.wav***		5,913	127
L4_3.wav		15,232	328	***L19_3.wav***		7,259	156
L5_1.wav	M	10,809	232	***L20_1.wav***	F	8,183	176
L5_2.wav		11,517	248	***L20_2.wav***		8,147	175
L5_3.wav		12,074	260	***L20_3.wav***		8,172	175
L6_1.wav	M	12,678	273	***L21_1.wav***	F	7,156	154
L6_2.wav		11,215	241	***L21_2.wav***		7,199	155
L6_3.wav		10,486	225	***L21_3.wav***		6,202	133
L7_1.wav	M	12,190	262	***L22_1.wav***	F	9,406	202
L7_2.wav		13,165	283	***L22_2.wav***		8,606	185
L7_3.wav		14,140	304	***L22_3.wav***		8,756	188
L8_1.wav	F	14,628	315	***L23_1.wav***	M	9,989	193
L8_2.wav		13,653	294	***L23_2.wav***		9,217	198
L8_3.wav		11,773	253	***L23_3.wav***		8,963	192
L9_1.wav	F	11,357	244	***L24_1.wav***	M	8,561	184
L9_2.wav		8,104	174	***L24_2.wav***		8,473	182
L9_3.wav		8,957	192	***L24_3.wav***		7,864	169
L10_1.wav	M	8,467	182	***L25_1.wav***	M	8,777	184
L10_2.wav		8,427	181	***L25_2.wav***		6,594	142
L10_3.wav		8,116	174	***L25_3.wav***		5,800	124
L11_1.wav	F	9,449	204	***L26_1.wav***	F	6,540	140
L11_2.wav		8,808	189	***L26_2.wav***		8,023	172
L11_3.wav		8,569	185	***L26_3.wav***		6,819	146
L12_1.wav	F	8,037	173	***L27_1.wav***	F	9,404	202
L12_2.wav		7,074	152	***L27_2.wav***		8,950	192
L12_3.wav		7,157	154	***L27_3.wav***		8,904	191
L13_1.wav	M	19,034	409	***L28_1.wav***	M	10,617	228
L13_2.wav		12,377	266	***L28_2.wav***		9,473	204
L13_3.wav		12,962	279	***L28_3.wav***		9,643	207
L14_1.wav	M	8,017	172	***L29_1.wav***	F	9,415	202
L14_2.wav		7,422	159	***L29_2.wav***		8,936	192
L14_3.wav		7,251	156	***L29_3.wav***		8,906	191
L15_1.wav	M	5,551	119	***L30_1.wav***	F	11,432	256
L15_2.wav		5,497	118	***L30_2.wav***		9,462	203
L15_3.wav		5,498	118	***L30_3.wav***		8,162	175

Foi utilizado o software *Cool Edit 2000*, para realizar a gravação das amostras de vozes, como mostrado na Figura 5.1 o ambiente gráfico durante a exibição do arquivo *L1_3.wav.*

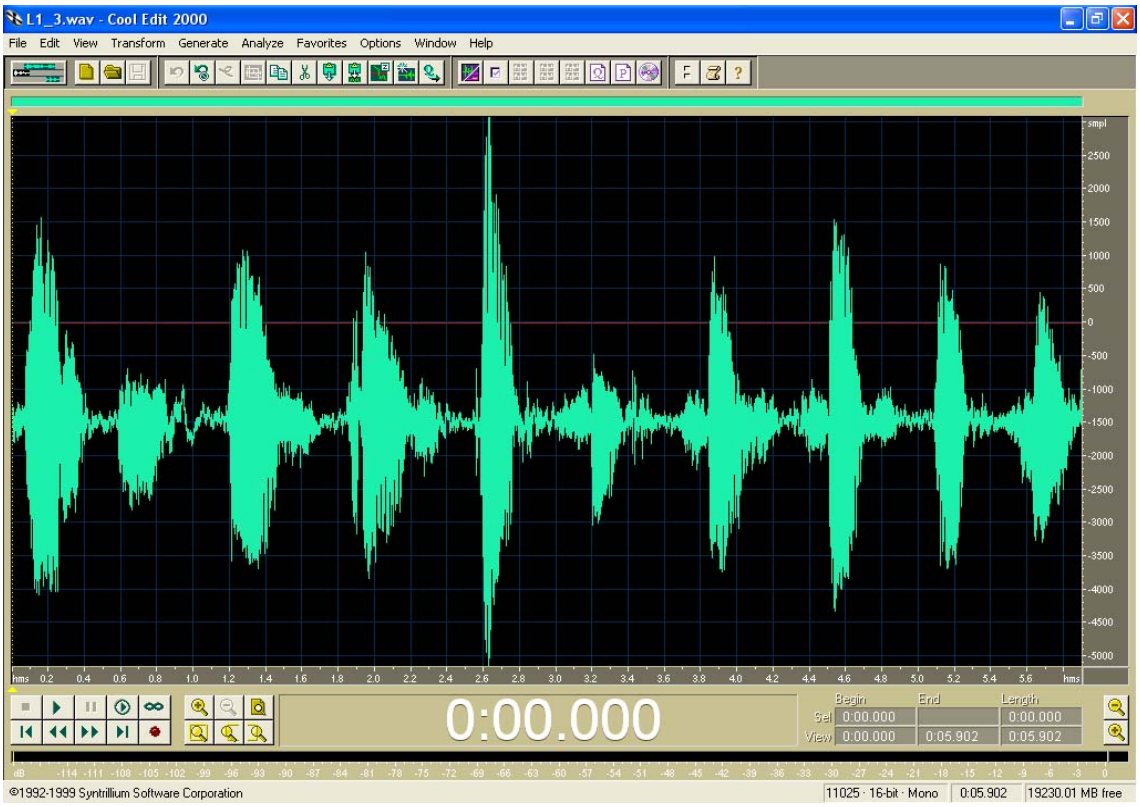


Figura 5.1 – Exibição do arquivo gravado *L1_3.wav* pela plataforma *Cool Edit 2000.*

5.3 - Estrutura do Gerador de Características

Para cada um dos arquivos do Banco de Dados foi feito a divisão em quadros de 256 amostras usando superposição de 128 amostras como pode ser visto na Figura 5.2., onde toma-se para efeito didático a forma de onda gerada pelo fonema /a/.

Os tipos de janelamentos mais usados são retangulares, Bartlet (triangular), Hanning, Hamming e Blackman. Mas a janela de Hamming é a mais comumente usada quando se trata de arquivos de áudio. Neste trabalho a janela de Hamming filtra o sinal,

devido ao fato de sua forma de onda possuir o lóbulo central muito maior que o secundário, mas sua principal função é eliminar as componentes de alta freqüência provenientes do janelamento.

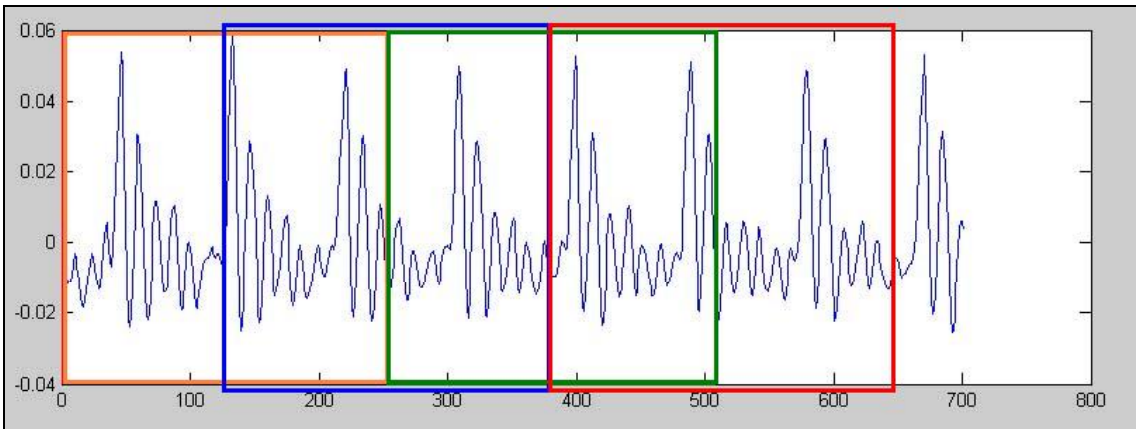


Figura 5.2 – Divisão em quadros de 256 amostras do fonema /a/ com superposição de 128 amostras.

Aplica-se então o janelamento de Hamming em cada janela de 256 amostras. A função janela de Hamming, $w_H(t)$ é obtida por:

$$w_H(t) = 0{,}54 + 0{,}46.\cos\left(\frac{2\pi t}{T}\right)$$

Na Figura 5.3. pode-se verificar a forma de onda de cada quadro após a aplicação do janelamento de Hamming no sinal visto na Figura 5.2.

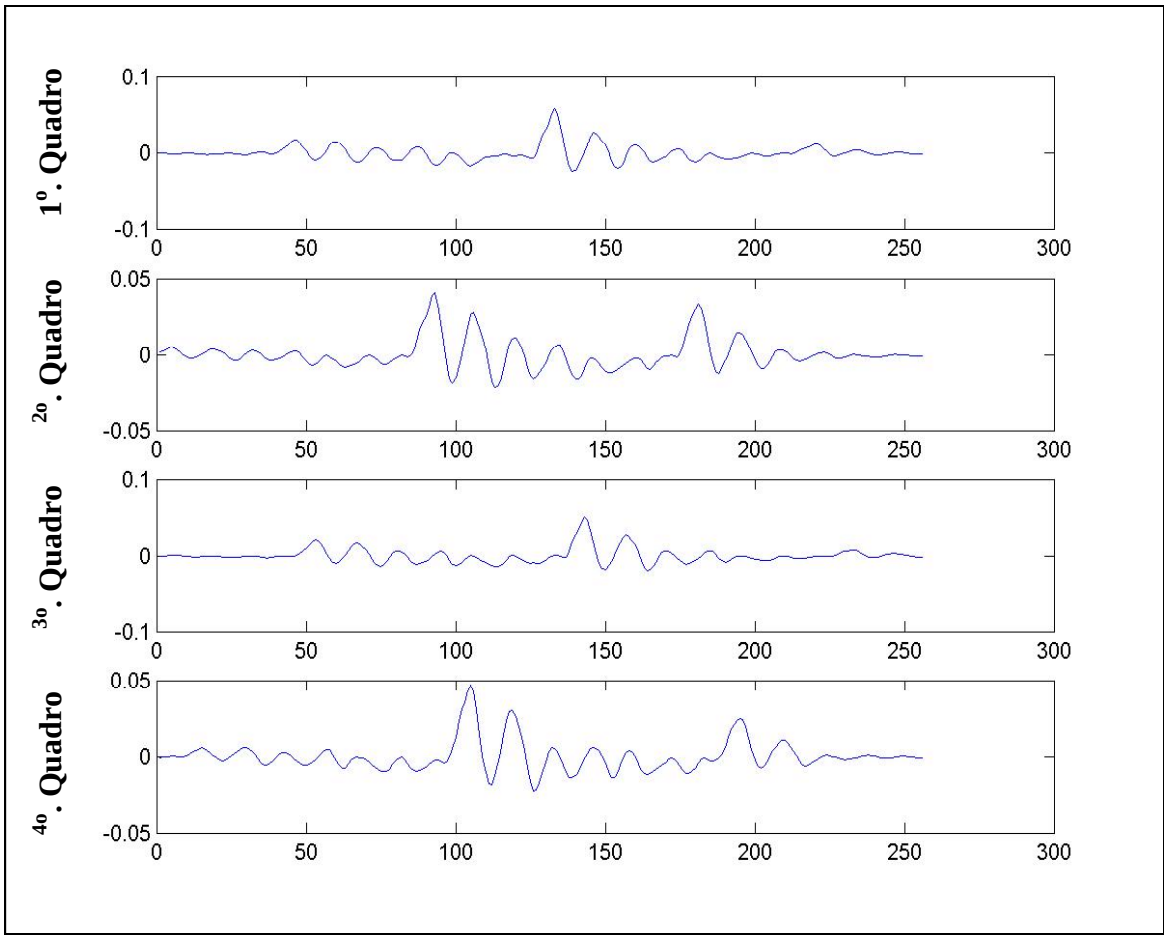


FIGURA 5.3 – Aplicação do janelamento de Hamming em cada quadro do sinal gerado pelo fonema /a/.

Para o banco de dados utilizado neste trabalho onde em média os arquivos possuem 10 min de gravação, o número de janelas é muito superior ao usado na ilustração acima, na Tabela 5.2 pode-se verificar isto.

Quando se manipula arquivos de áudio para efeito de processamento é necessário transformar essas amostras em um conjunto menor de características que tenha uma boa representação do sinal original, assim é necessário usarmos uma medida de características.

TABELA 5.2 – Tabela da contagem de janelas para cada um dos arquivos do banco.

	Número de Janelas						
Locutor	**Arq. 1**	**Arq. 2**	**Arq. 3**	**Locutor**	**Arq. 1**	**Arq. 2**	**Arq. 3**
L1	*858*	*994*	*508*	***L16***	*534*	*569*	*563*
L2	*1041*	*1075*	*1074*	***L17***	*565*	*481*	*515*
L3	*848*	*854*	*575*	***L18***	*588*	*535*	*522*
L4	*1365*	*1215*	*1311*	***L19***	*507*	*509*	*625*
L5	*931*	*991*	*1039*	***L20***	*704*	*701*	*703*
L6	*1091*	*965*	*903*	***L21***	*616*	*620*	*534*
L7	*1049*	*1133*	*1217*	***L22***	*810*	*741*	*754*
L8	*1259*	*1175*	*1014*	***L23***	*774*	*793*	*769*
L9	*978*	*698*	*771*	***L24***	*737*	*729*	*677*
L10	*729*	*725*	*699*	***L25***	*755*	*567*	*499*
L11	*818*	*758*	*740*	***L26***	*563*	*691*	*587*
L12	*692*	*609*	*616*	***L27***	*809*	*770*	*766*
L13	*1639*	*1066*	*1116*	***L28***	*914*	*815*	*830*
L14	*690*	*639*	*624*	***L29***	*810*	*769*	*767*
L15	*478*	*473*	*473*	***L30***	*984*	*814*	*703*

Utiliza-se *Linear Predictive Coding (LPC)* como medida das características, após a divisão em quadros e a filtragem do sinal calcula-se então os 12 coeficientes LPC para cada uma das janelas, gerando uma matriz *n* x 12, onde *n* varia de acordo com o número de janelas.

5.4 - Geração dos ‘Codebooks’

Na Tabela 5.1 pode-se verificar que raramente um sinal coincide com o outro, em tamanho, com isso houve também uma variação muito grande entre o número de quadros da Tabela 5.2, superior a 170 % entre o menor e o maior arquivo amostrado. Assim é necessário uma solução para adaptar os valores que são analisados com os de parâmetros e essa é a utilização da *Dynamic Time Warping* (DTW), que faz uma

compensação nesta variação no domínio do tempo. Após calcular os coeficientes do LPC já pode-se fazer uma análise da voz, sabendo-se que os vetores formados por eles já são os vetores característicos, como foi feito em [8].

Porém uma outra solução foi adotada, que seria a geração de *codebooks* seguindo o algoritmo LBG como mostra o capítulo IV deste trabalho. Obtem-se então uma matriz, que representa os vetores característicos ou centróides, de ordem m x 12, onde m varia de acordo com a quantidade de centróides que foi utilizada. Assim independente do número de janelas, a dimensão das matrizes que são analisadas é sempre a mesma, desde que seja constante o número de centróides.

Foi associado então, para cada um dos locutores pertencentes ao banco de dados, 3 *codebooks* para um mesmo número de centróides. Neste trabalho foram gerados *codebooks* com 4, 8, 16 e centróides para cada um dos arquivos do banco de dados, ou seja, 9 *codebooks* para cada um dos locutores totalizando 270 *codebooks*.

5.5 - Método de Treinamento

Após a criação dos *codebooks* é necessário criar os "*speaker models"*, ou modelos dos locutores, este que funcionam como "peso" no sistema.

Primeiramente tomar-se os arquivos de cada locutor para criarmos esses modelos, para efeito de esclarecimento esses arquivos foram *LS_1.wav* e *LS_2.wav* (para referência), *LS_3.wav* (para análise) onde $S=\{1,2,3,...,30\}$, isto é, varia de acordo com os locutores em questão. Tomaremos os seus respectivos *codebooks* e construiremos a polinomial para cada um dos arquivos.

Feito isso obtem-se uma matriz, P para cada um dos *codebooks*, de ordem k x nc, onde k é o grau da polinomial, ou seja, é determinado pelo grau de $p(x)$ e nc é número

de centróides utilizado. A expansão da polinomial para uma vetor característico de dimensão 12 pode ser feito da seguinte forma, a função *p(x)* polinomial recebe o valor 1 (um) na primeira coordenada e segue com os produtos de cada uma das coordenadas do vetor característico , ou seja,

$$p(x) = [1 \quad x_1 \quad x_2 \quad \ldots \quad x_{12} \quad x_1^2 \quad x_1x_2 \quad x_1x_3 \quad \ldots \quad x_{12}^2]^t$$

Como vimos no capítulo IV deste trabalho, ou ainda como foi feito por *Wan* e *Renals* em [30].

Assim utilizando-se 12 coeficientes do LPC e *nc* = 16 , a matriz *P* é da ordem *91x16*, visto que o número de elementos de *p(x)* é obtido a Equação (4.11) da seção 4.4.2 do capítulo IV deste trabalho.

O próximo passo é determinarmos a matriz M_{loc}, apartir da concatenação das matrizes *P*. Quando for utilizado 2 arquivos de parâmetro para cada locutor, M_{loc} éobtida da seguinte forma:

$$M_{loc} = \begin{bmatrix} P_{11}^t \\ P_{12}^t \\ \cdot \\ \cdot \\ \cdot \\ P_{N1}^t \\ P_{N2}^t \end{bmatrix},$$

onde *N* é o número de locutores que serão analizados.

Calcula-se agora a matriz uma matriz R_{ij} , obtida por:

$$R_{ij} = M_{ij}^{t} * M_{ij},$$

onde *i* varia de acordo com o número de locutores parâmetros, e *j* com o número de arquivos utilizados de cada locutor para estabelecer os parâmetros. Feito isso obtem-se a matriz *R* obtida por:

$$R = \sum_{i=1}^{L} \sum_{j=1}^{2} R_{ij}$$

É preciso agora obter as matrizes O_N, que tem dimensão (*N.nc)* x 1, que é construída da seguinte forma, matriz coluna formadas por 1 (um), nas coordenadas que correspondem aos locutor em questão e 0(zero) nas demais, assim por exemplo vejamos o seguinte exemplo:

Seja um sistema de reconhecimento formado por 5 locutores, utilizando apenas 1 arquivo de parâmetro para cada um dos locutores em questão e um codebook com apenas 2 centróides, as matrizes O_i's são as seguintes:

$$O_1 = [1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$O_2 = [0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$O_3 = [0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$O_4 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0]^T$$

$$O_5 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1]^T$$

Generaliza-se então para *N* locutores e obtém-se o seguinte conjunto de matrizes O_i's:

$$O_1 = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} , \ldots , \quad O_N = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$$

onde a matriz O_1 está associada ao locutor 1, assim sucessivamente até a matriz O_N que está associada ao locutor N.

Assim determina-se os modelos dos w_N 's locutores, da seguinte forma:

$$R.w_N = M^t.O_N$$

Após encontrar w_N para cada uma dos locutores calcula-se a média do produto desses por $p(x_i)$ obtendo-se o *score*, s_i, ou seja:

$$s_i = \frac{1}{N} \sum_{j=1}^{N} w_i^t \, p(x_j)$$

onde: ***N*** é o número de locutores que foram utilizados na análise;

j varia de acordo com o locutor parâmetro, e;

i varia de acordo com o locutor teste.

Encontrados os $\boldsymbol{s_i}$'s, seguiremos com a análise de forma diferente a realizar o verificação ou a identificação como veremos no próximo item.

5.6- Verificação e Identificação

Para a verificação do locutor, ou seja, a tomada de decisão do sistema de aceitar ou rejeitar um determinado locutor ou um grupo deles, é feita pela avaliação do comportamento dos s_i's por um limiar (l). Assim sempre que $s_i < l$ rejeita-se a solicitação, caso contrário aceita-se.

No caso da identificação, o processo é mais simples. O solicitante é identificado como sendo o maior valor de s encontrado, porém é dotado de uma maior número de cálculos. Pois ao invés de determinar-se N matrizes O, como é feito nesse processo na verificação só é necessário uma. E ainda os pesos necessários para o cálculo do score são feitos para cada locutor, enquanto na verificação é necessário apenas um [7].

5.6.1- A verificação

Para aplicar o algoritmo de verificação deve-se primeiramente determinar os membros com verificação positiva, representados pelo índice *spk*, e quais tem verificação negativa, representada pelo índice *imp*. Feito isso toma-se dois arquivos de cada um dos locutores aptos (*spk*) e inaptos (*imp*) para criar o *w*, como foi descrito anteriormente. Seguido pelo cálculo dos s_i's.

Feito isso determina-se o limiar da verificação, $s_{\lim}$, obtido por:

$$s_{\lim} = \min\{s_{spk}\}$$

Assim acompanham-se o desempenho do sistema quando varia-se o número de centróides como mostrado nas Figuras 5.4, 5.5 e 5.6 para *4*, *8* e *12* centróides,

respectivamente. Foi adotado para estes testes um número fixo de 10 locutores aptos e 20 inaptos. Nessas figuras a posição dos usuários em relação ao limiar deve ser a seguinte: usuários aptos devem estar acima do limiar e inaptos abaixo.

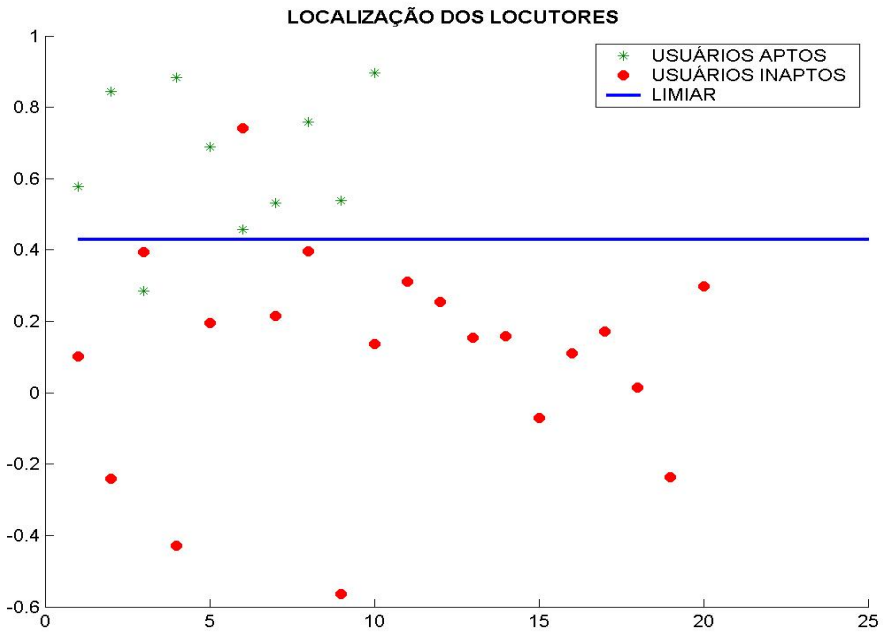


Figura 5.4 – Sistema utilizando 10 locutores aptos, 20 inaptos, com 4 centróides.

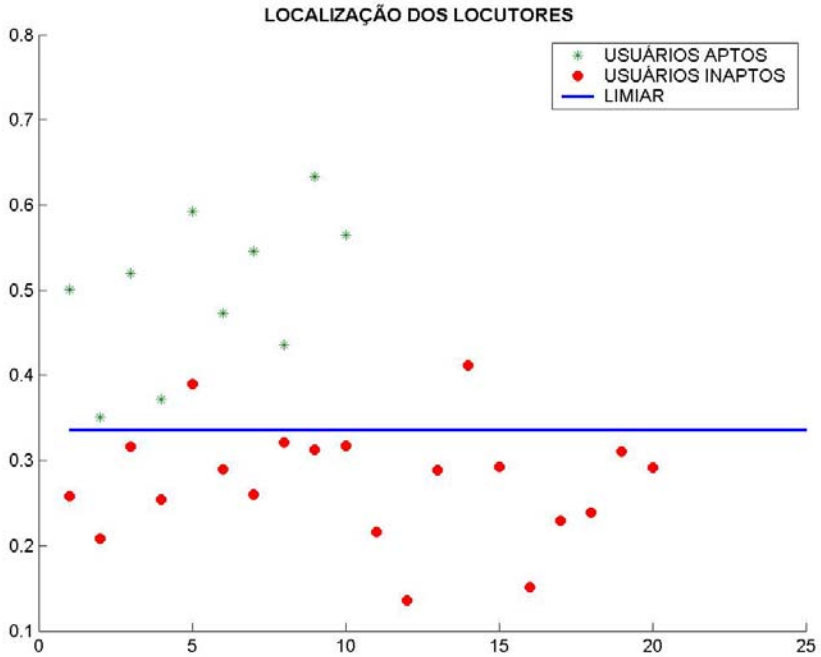


Figura 5.5 – Sistema utilizando 10 locutores aptos, 20 inaptos, com 8 centróides.

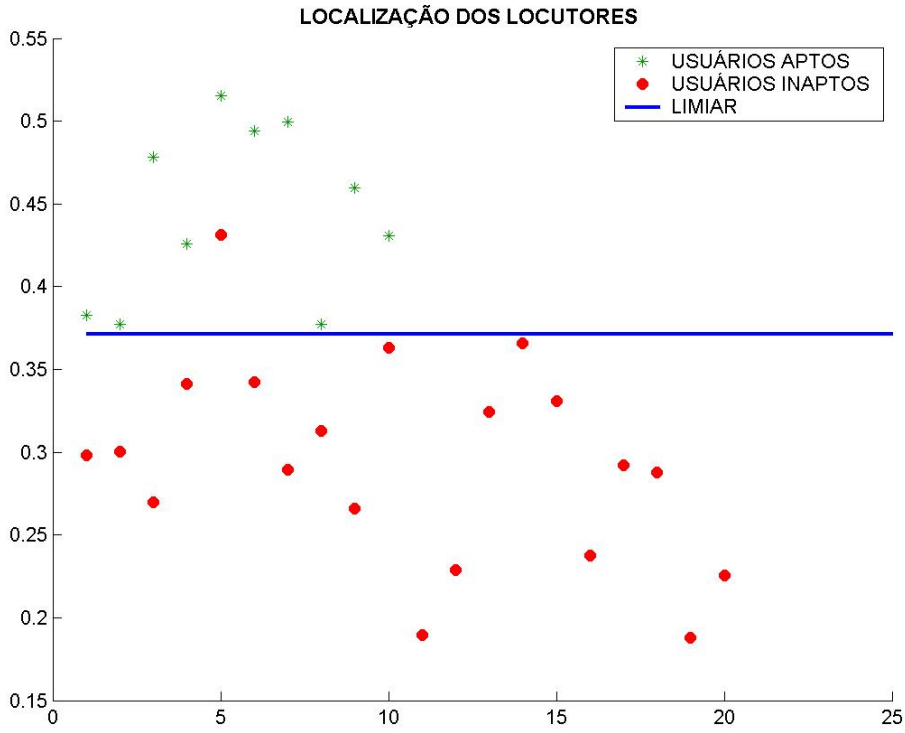


Figura 5.6 – Sistema utilizando 10 locutores aptos, 20 inaptos e com16 centróides.

Observe na Figura 5.4 que ocorreram 2 tipos de erros, um na classificação de aptidão, chamado erro de aptidão (e_A), que ocorre quando o locutor está inapto mas é classificado como apto, e outro na inaptidão, chamado erro de inaptidão (e_I) , que ocorre quando o locutor está apto e é classificado como inapto. Quando aumenta-se o número de centróides os erros são apenas de aptidão e decresce a medida que aumenta-se o número de centróides como pode ser visto nas Figuras 5.5 e 5.6. Na Tabela 5.3 o comportamento pode ser melhor analisado.

Tabela 5.3 – Ocorrência do erro com variação do número de centróides.

Nº. de centroides	**Limiar**	**Rejeição Máxima**	**Aceitação Mínima**	e_A / **Ináptos**	e_I / **Aptos**	e_T / **Total**	% e_A	% e_I	% e_T
4	*0,4300*	*0,7405*	*0,4573*	*1/20*	*1/10*	*2/30*	*5*	*10*	*6,67*
8	*0,3053*	*0,4120*	*0,3511*	*2/20*	*0/10*	*2/30*	*10*	*0*	*6,67*
16	*0,3413*	*0,3657*	*0,3771*	*1/20*	*0/20*	*1/30*	*5*	*0*	*3,34*

5.6.2- A identificação

Toma-se agora dois arquivos de cada locutor que deseja-se reconhecer para estabelecer os parâmetros. Neste algoritmo tem-se que calcular um w_j para cada locutor que é reconhecido, como já foi dito. Apresenta-se a seguir os resultados encontrados usando o algoritmo de reconhecimento.

Primeiramente veja o que acontece, quando gera-se um codebook com 16 centróides e varia-se o número de locutores.

O gráfico na Figura 5.7 apresentam as variação de $\boldsymbol{s_{ij}}$ reconhecendo 5 locutores, sendo 2 do sexo feminino e 3 do sexo masculino. Em que s_{ij}, é obtido pela média do produto de $p(x_i).w_j$, como foi descrito anteriormente, sendo assim o s_{ij} de maior nível classificará o arquivo origem de $p(x_i)$ como sendo da mesma classe que o arquivo de origem de w_j . Assim fica como reconhecimento verdadeiro os s_{ij}, onde $i = j$ possuem o maior nível.

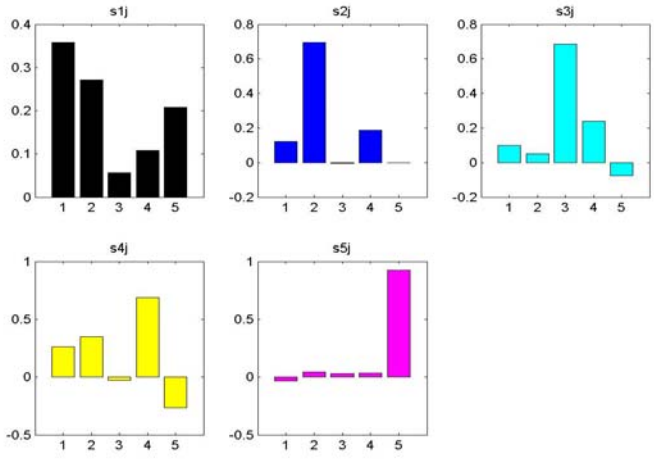


Figura 5.7- Score dos locutores 1 ao 5 de 5, com 16 centróides.

Note que, até mesmo pelo fato de se tratar de um número pequeno de locutores, o erro é de 0%.

Na Figura 5.8 a seguir, pode-se verificar o comportamento do sistema no reconhecimento de 10 locutores em que 4 são do sexo feminino e 6 são do sexo masculino.

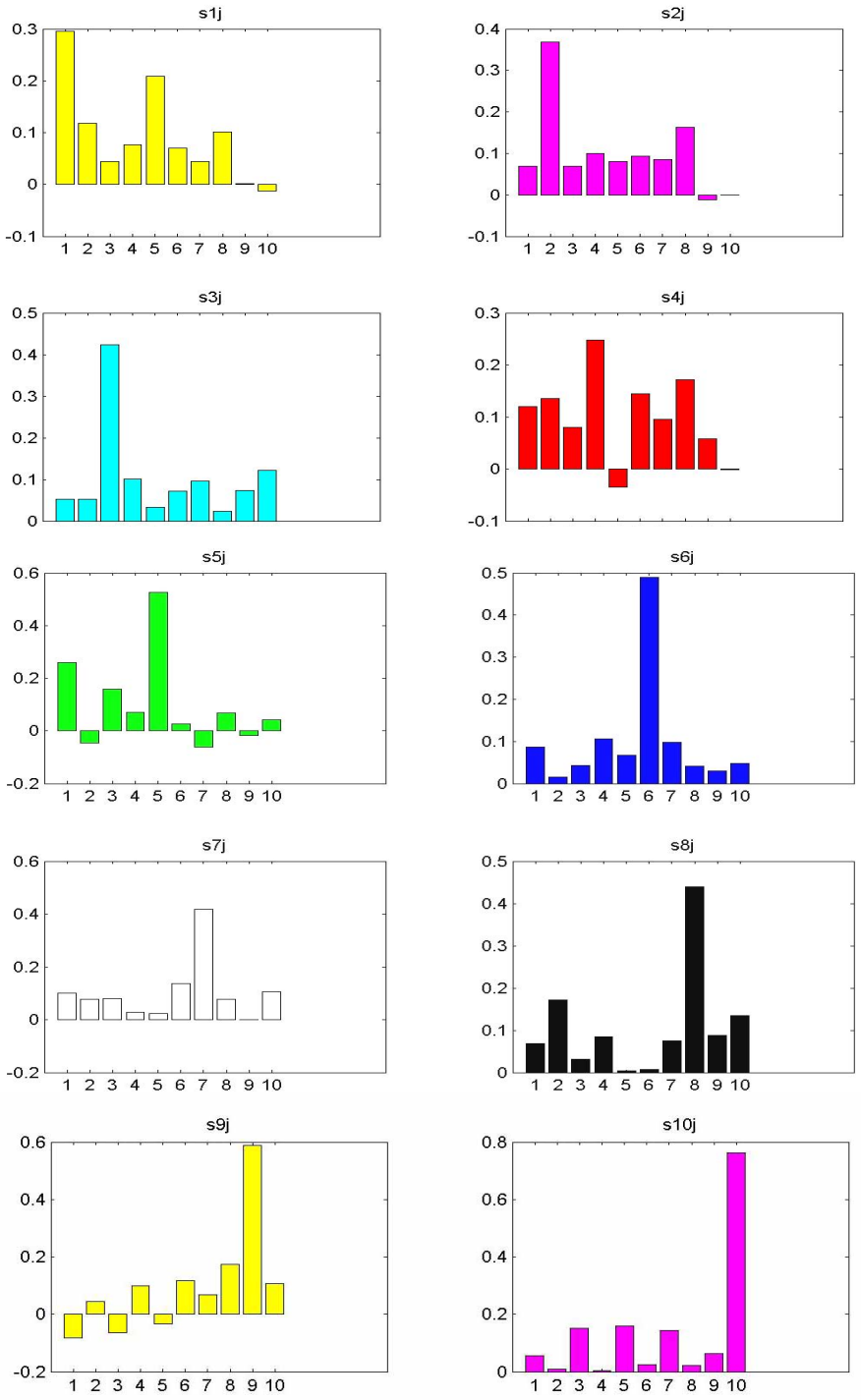


Figura 5.8 – Score dos locutores 1 ao 10 de 10, com 16 centróides.

Agora veja o comportamento do sistema quando é desejado que ele identifique 15 locutores, Figuras 5.9.

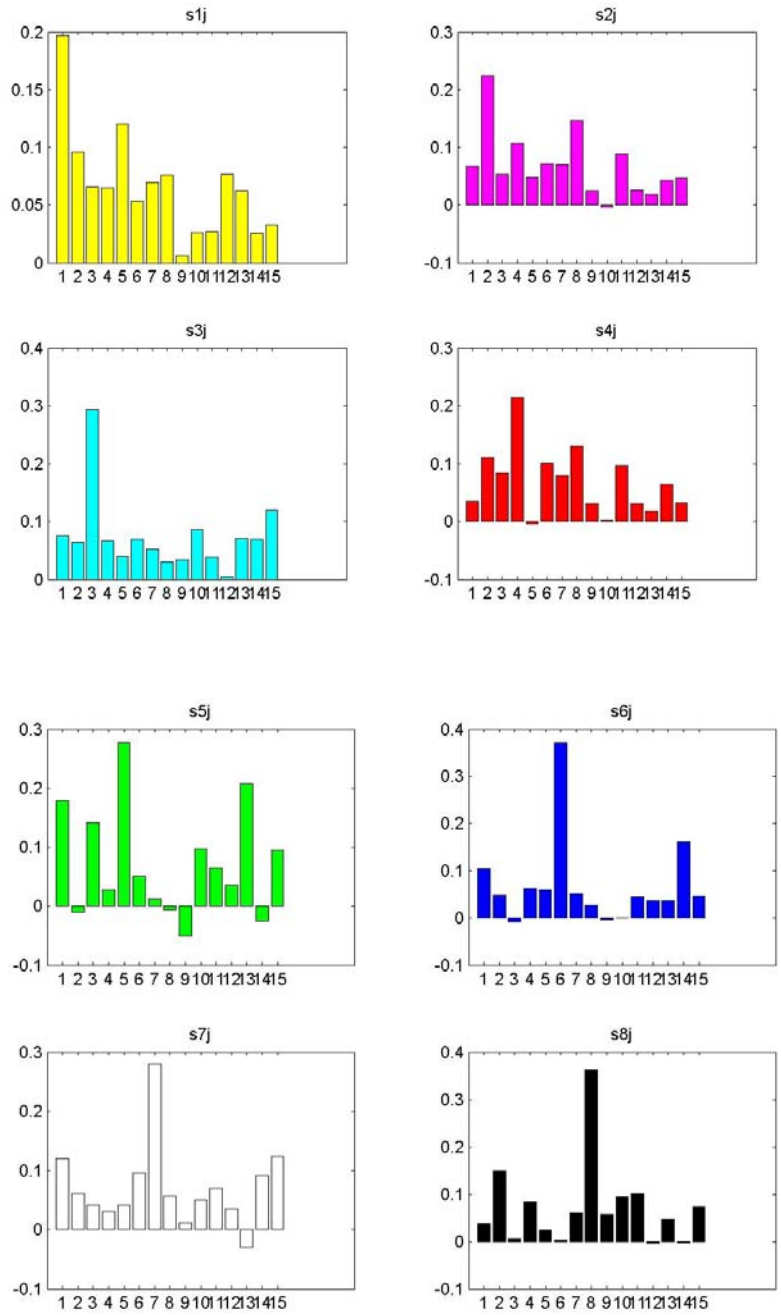


Figura 5.9.1 – Score dos locutores 1 ao 8 de 15, com 16 centróides.

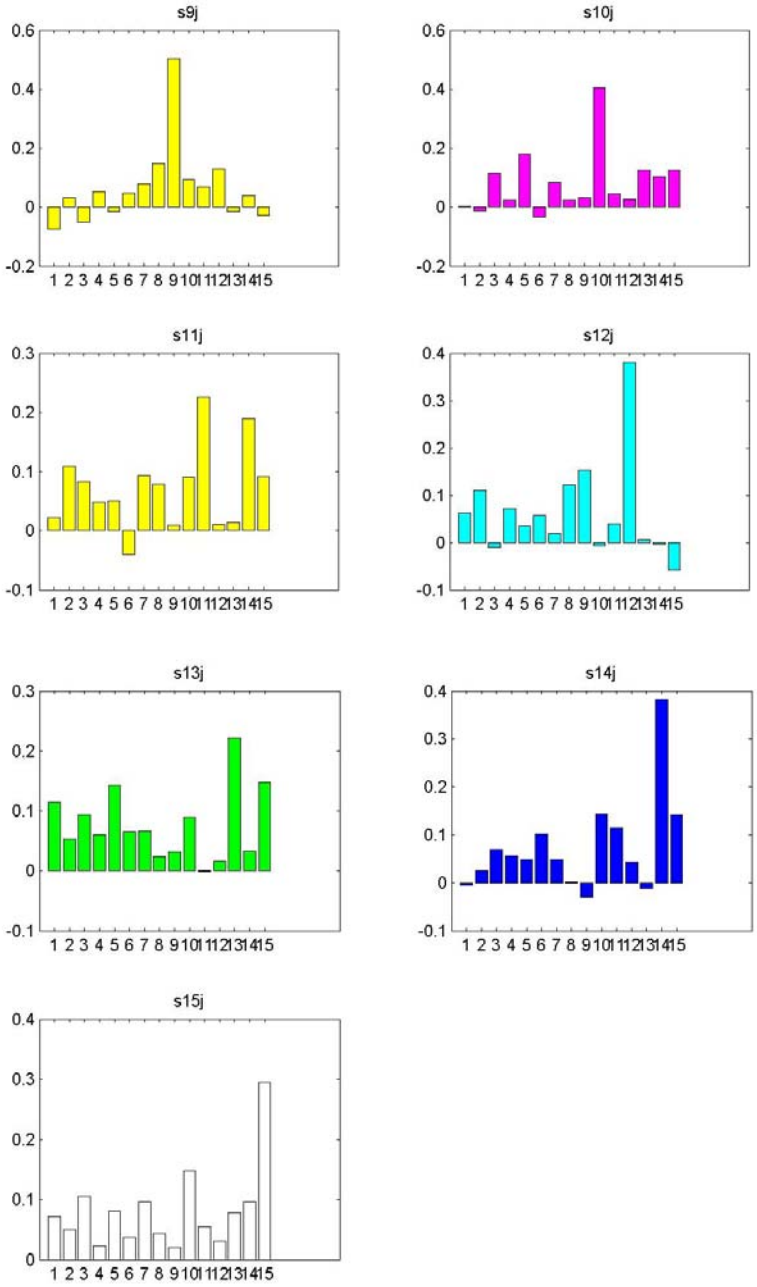


Figura 5.9.2 – Score dos locutores 9 ao 15 de 15, com 16 centróides.

Pode-se perceber também a facilidade com que o sistema reconheceu os 15 locutores com eficiência de 100%.

Agora introduzindo-se mais 5 locutores no sistema e pode-se perceber a ocorrência de erro na identificação de 1 locutor, como mostrado na Figura 5.10.3, em s_{18j},

em que a maior incidência deste está em $j = 19$. Portanto erro de 5%. Nessa configuração foram usadas 10 locutores do sexo feminino e 10 do sexo masculino. Vejamos o desempenho do sistema no reconhecimento de 20 locutores nas Figuras 5.10.

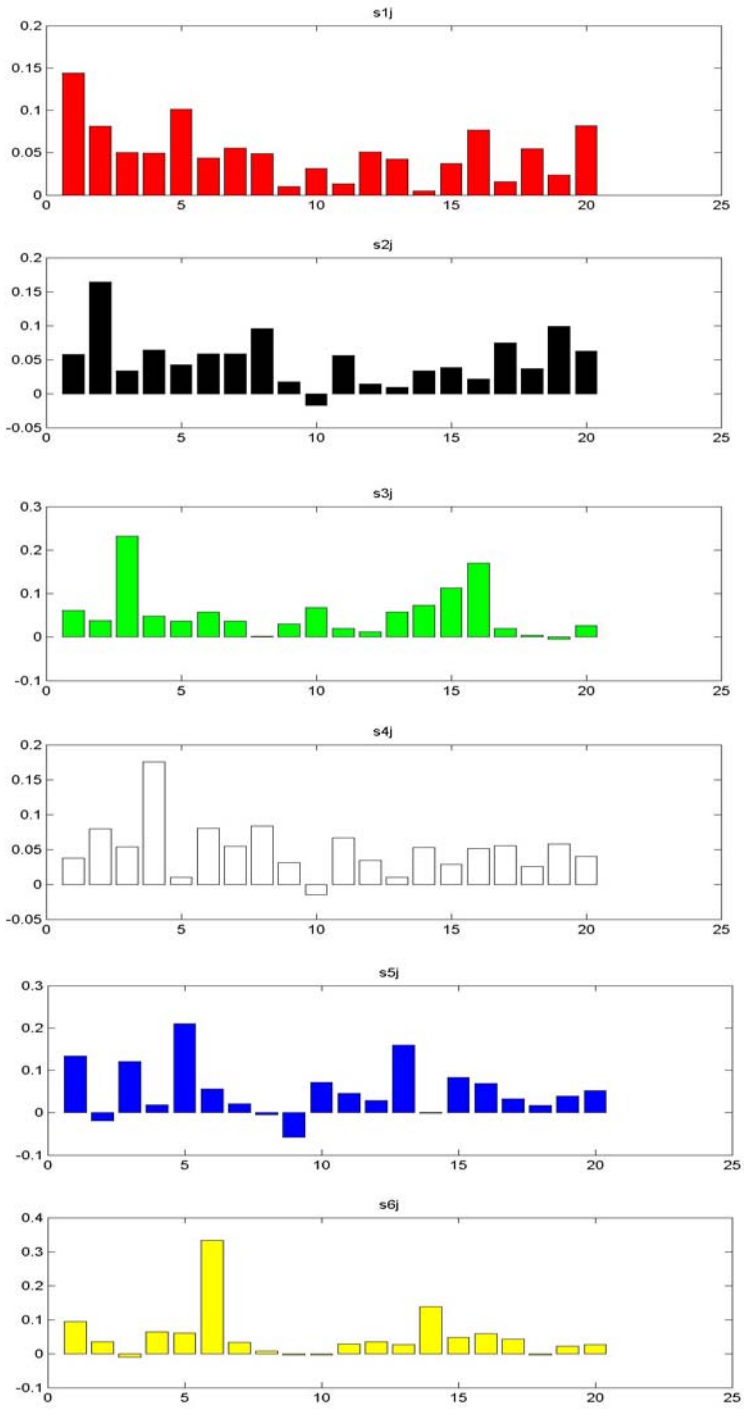


Figura 5.10.1 – Score dos locutores 1 ao 6 de 20, com 16 centróides

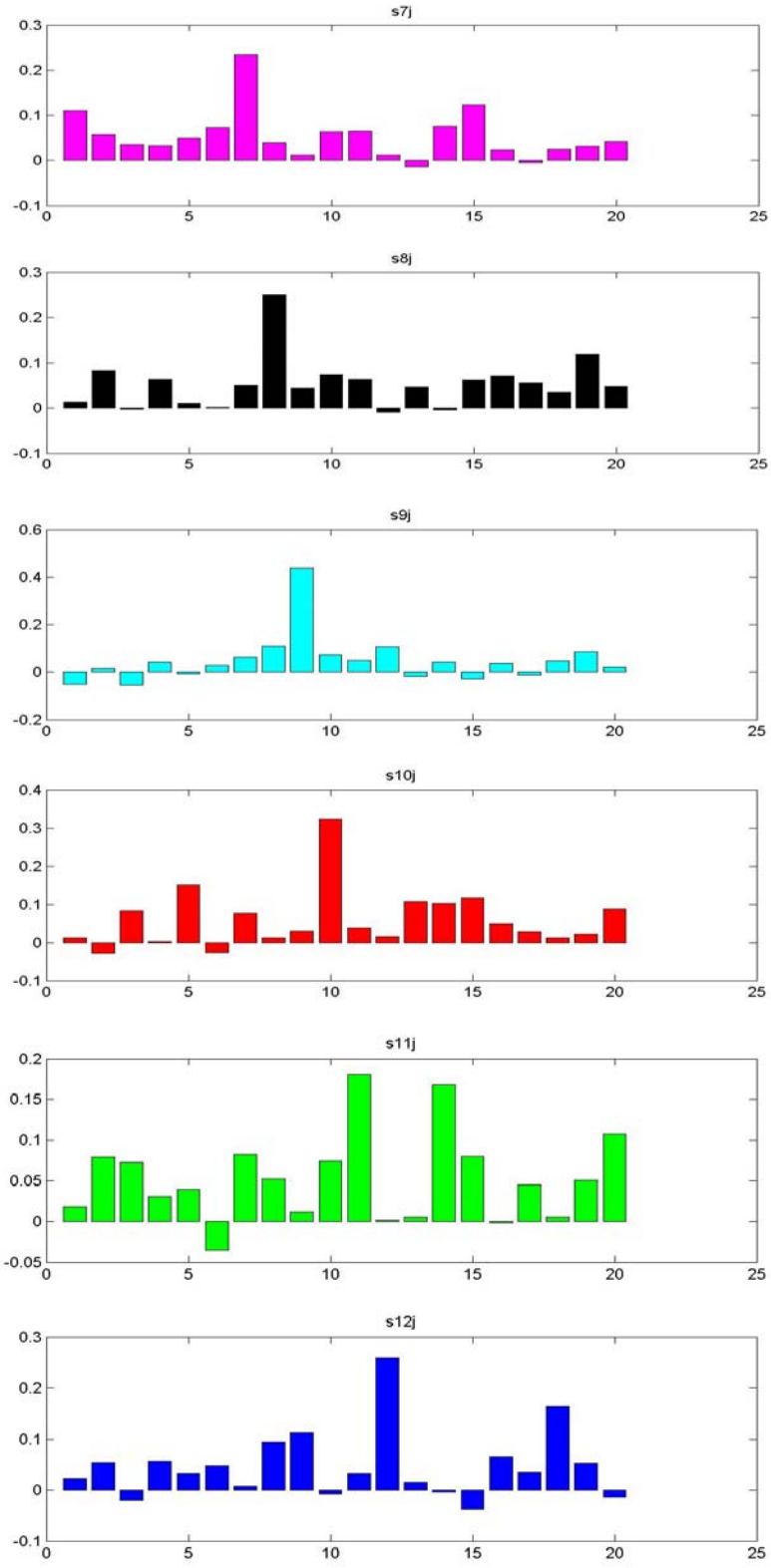


Figura 5.10.2 – Score dos locutores 7 ao 12 de 20, com 16 centróides

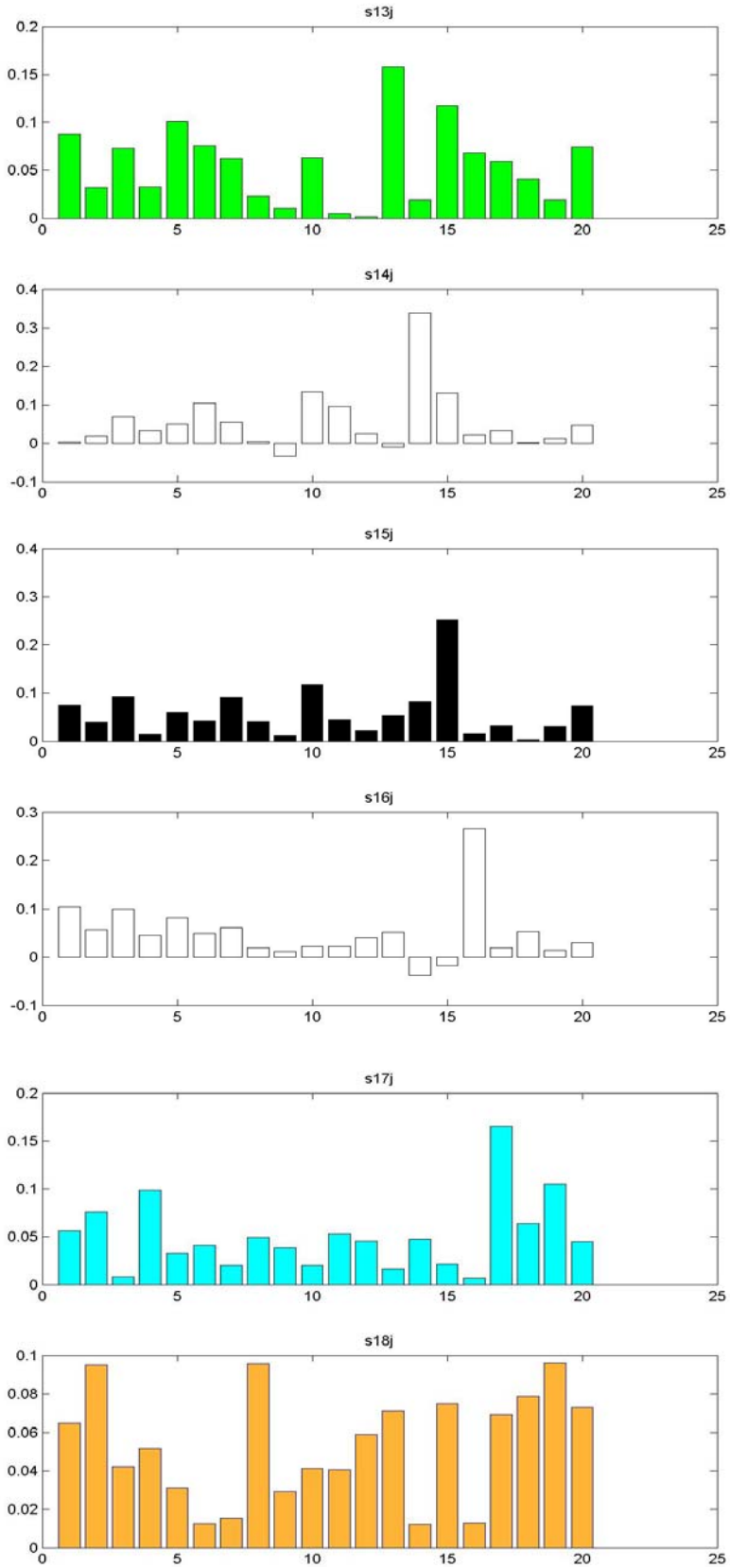


Figura 5.10.3 – Score dos locutores 13 ao 18 de 20, com 16 centróides

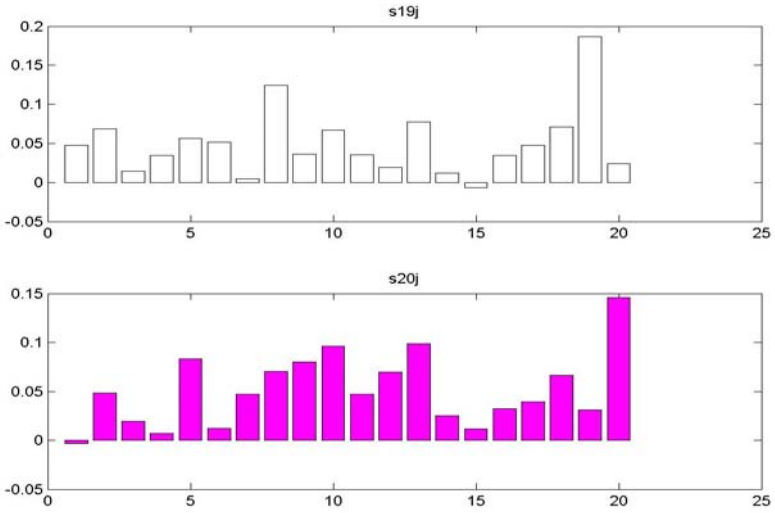


Figura 5.10.4 – Score dos locutores 19 e 20 de 20, com 16 centróides

Posteriormente ao aumentarmos o número de locutores do sistema, de 20 para 25 locutores sendo 12 do sexo feminino e 13 do sexo masculino, o erro cai para 4%. Isto pode ser visto nas Figuras 5.11 e a incidência de erro na Figura 5.11.4, novamente em s_{18j} , em que verifica-se o máximo em $j = 19$.

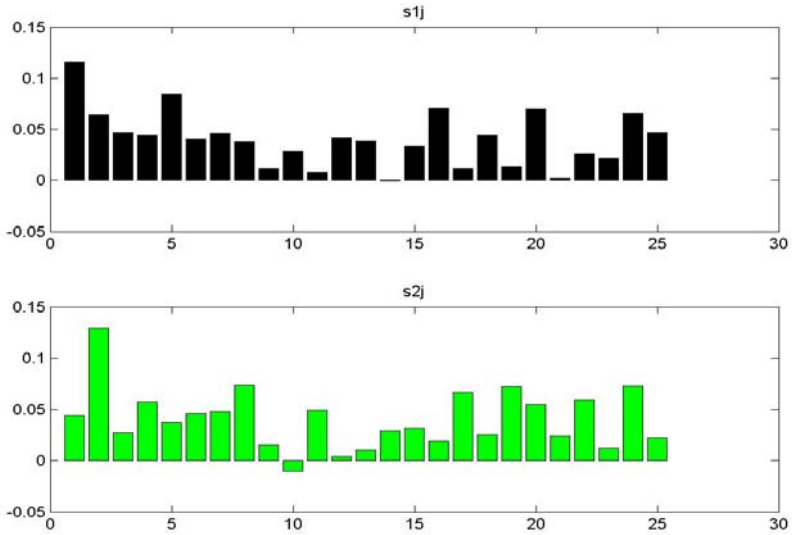


FIGURA 5.11.1 – Score dos locutores 1 e 2 de 25, com 16 centróides.

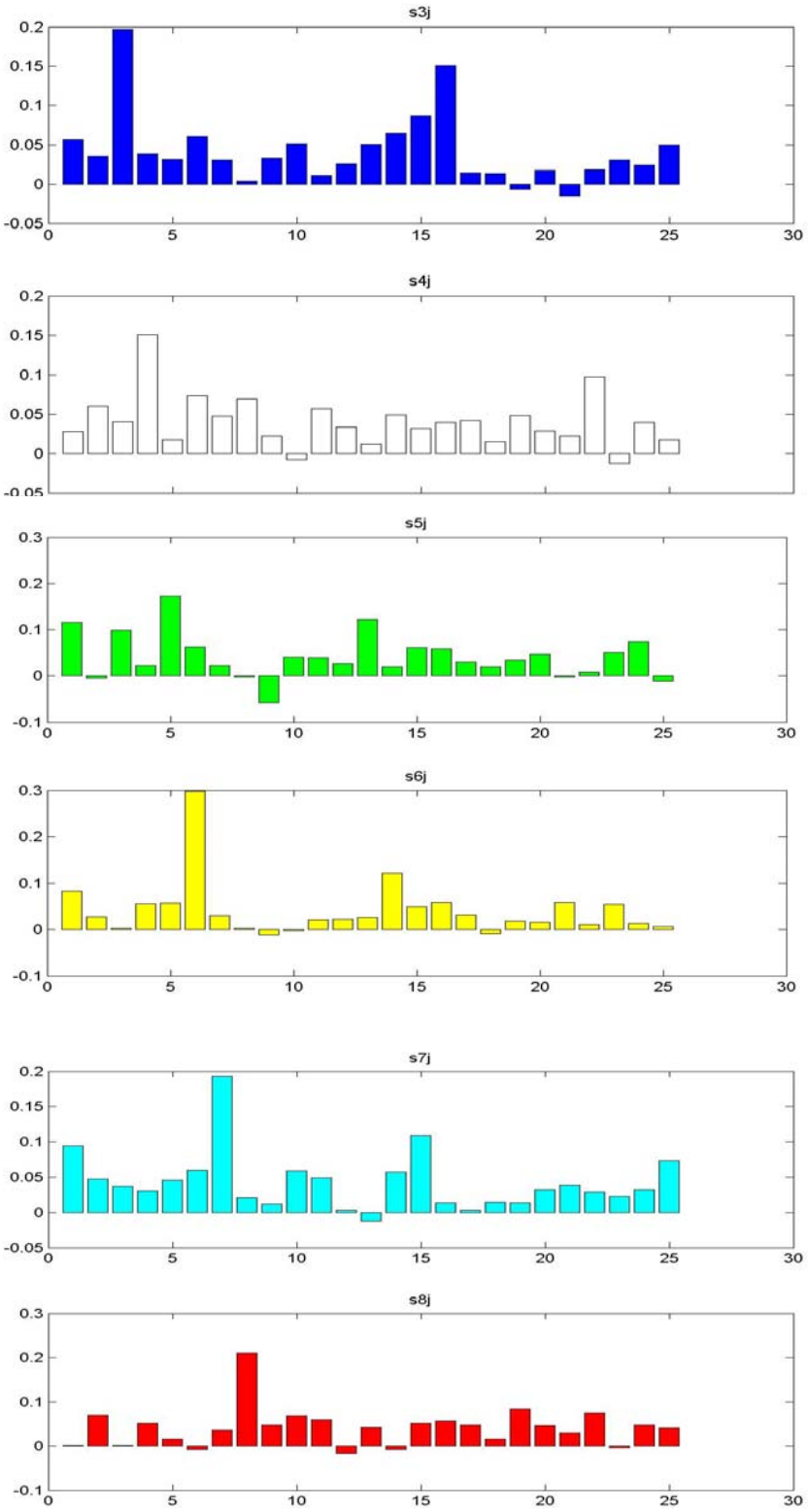


Figura 5.11.2 – Score dos locutores 3 ao 8 de 25, com 16 centróides.

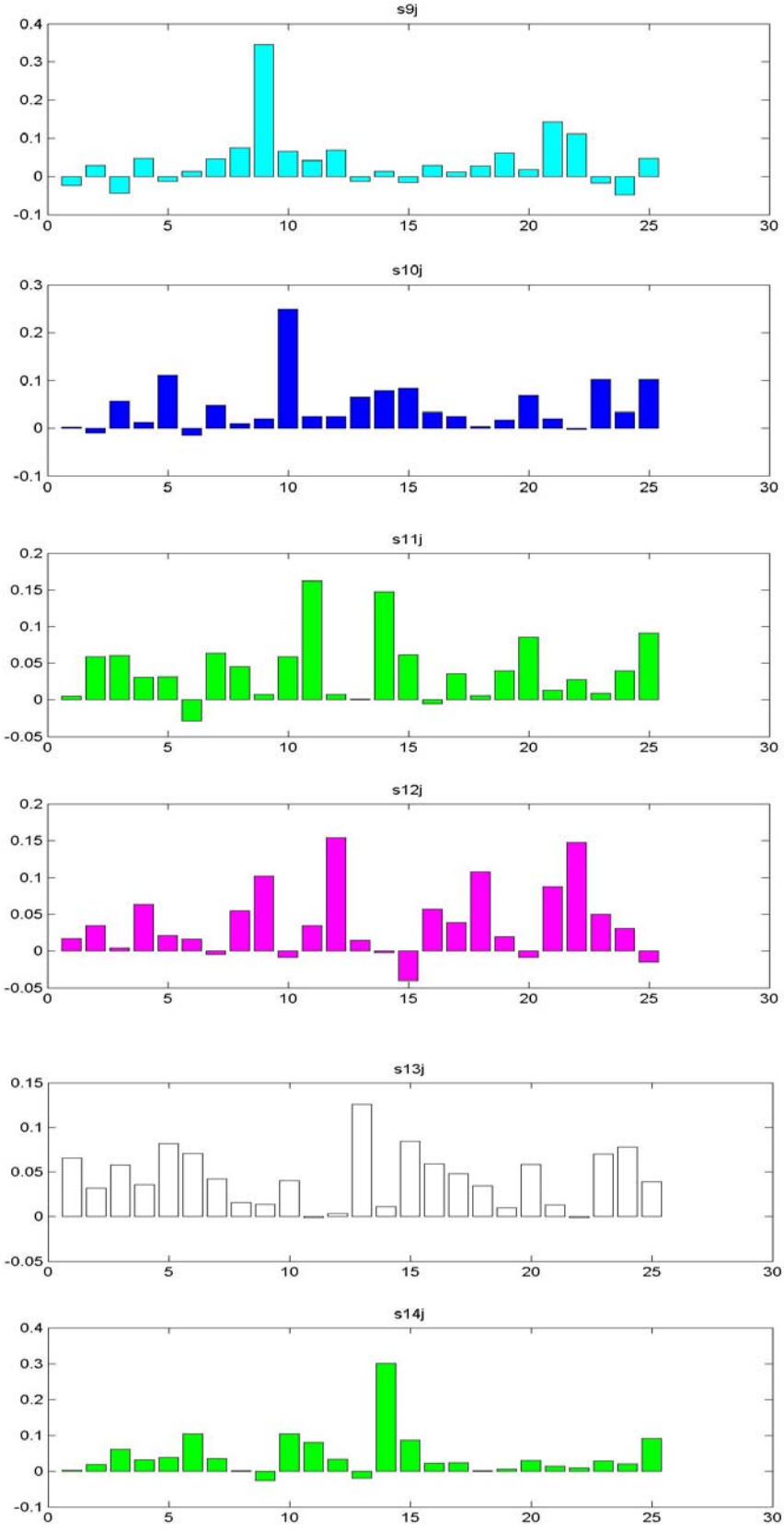


Figura 5.11.3 – Score dos locutores 9 ao 14 de 25, com 16 centróides.

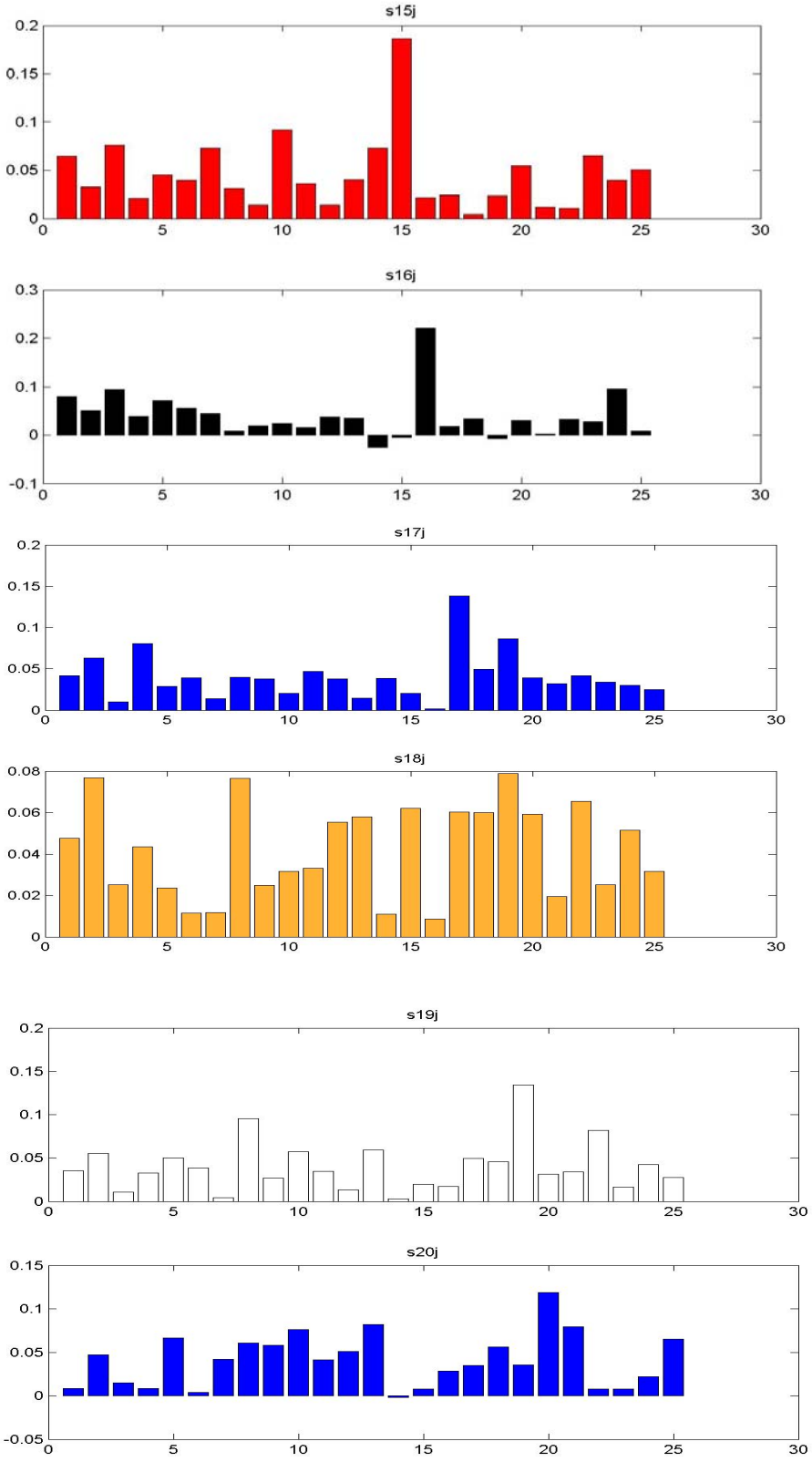


Figura 5.11.4 – Score dos locutores 15 ao 20 de 25, com 16 centróides.

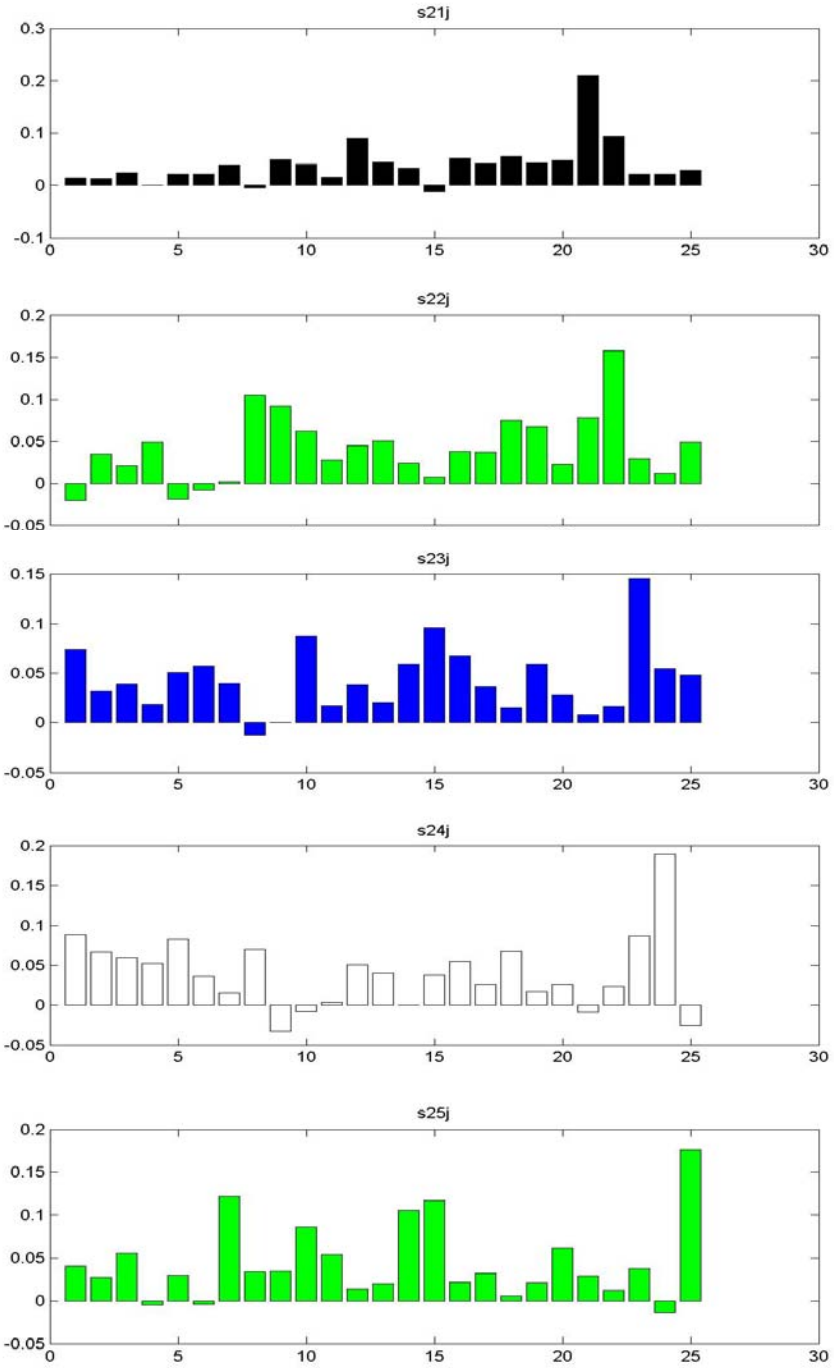


Figura 5.11.5 – Score do locutor 21 a 25 de 25, com 16 centróides.

Ao aumentar o número de locutores a serem reconhecidos para 30, novamente ocorreu a redução do erro, para *3,33%* , já que s_{ij} com $i = 18$, teve seu arquivo reconhecido como sendo de $j = 26$, isso pode ser visto na Figura 5.12.9. Pode se ver o comportamento

para esta configuração, 30 locutores sendo 16 do sexo feminino e 14 do sexo masculino, nas Figuras 5.12.

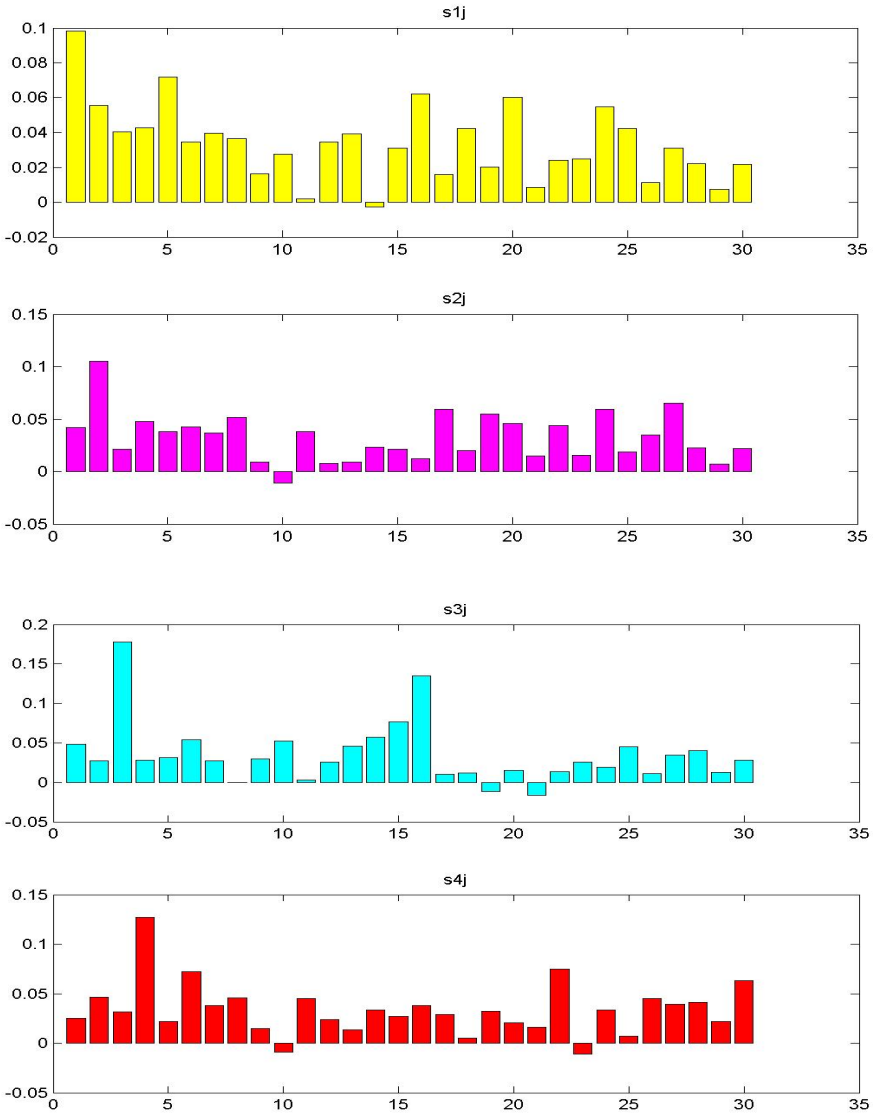


Figura 5.12.1 – Score dos locutores 1 ao 4 de 30, com 16 centróides.

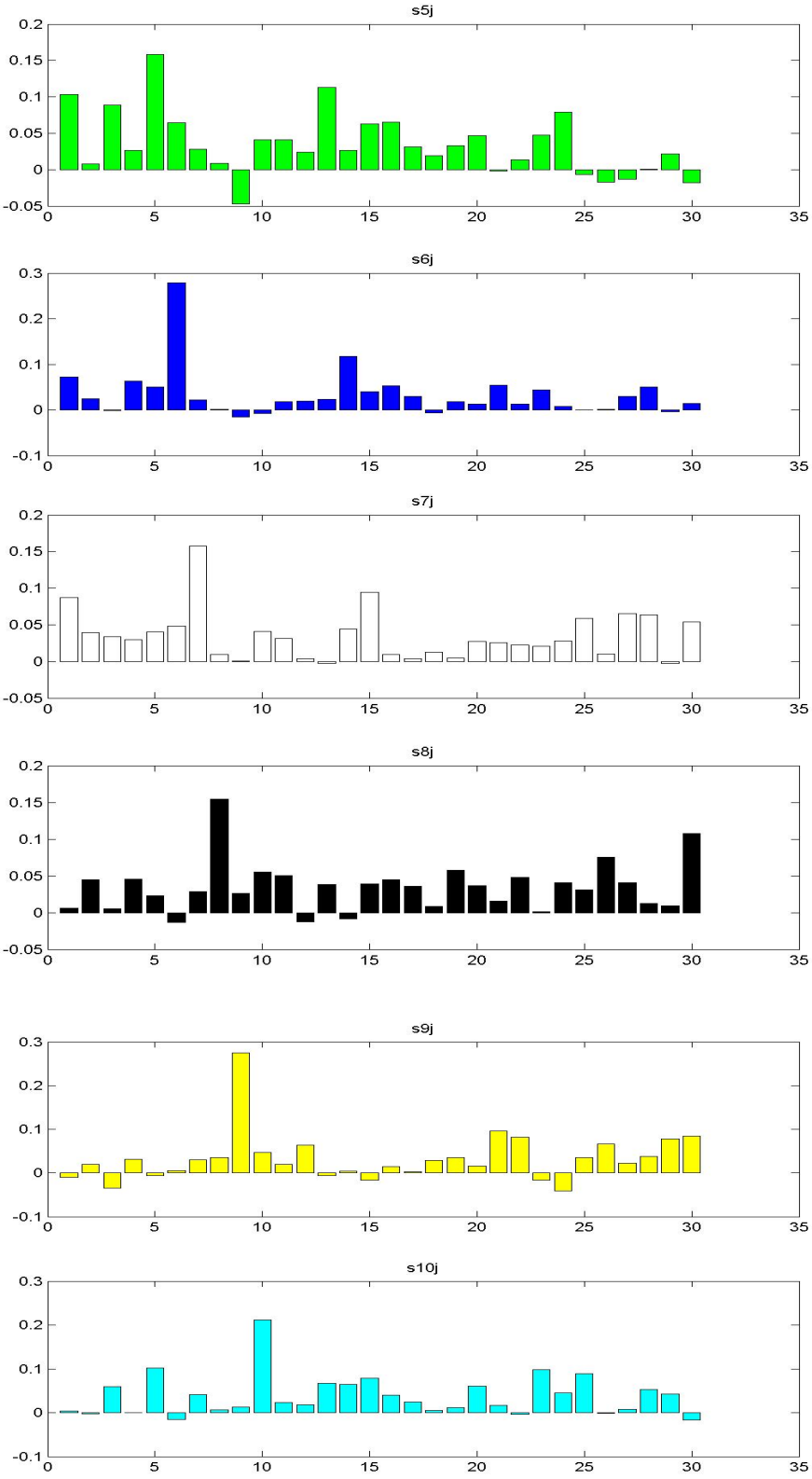


Figura 5.12.2 – Score dos locutores 5 ao 10 de 30, com 16 centróides.

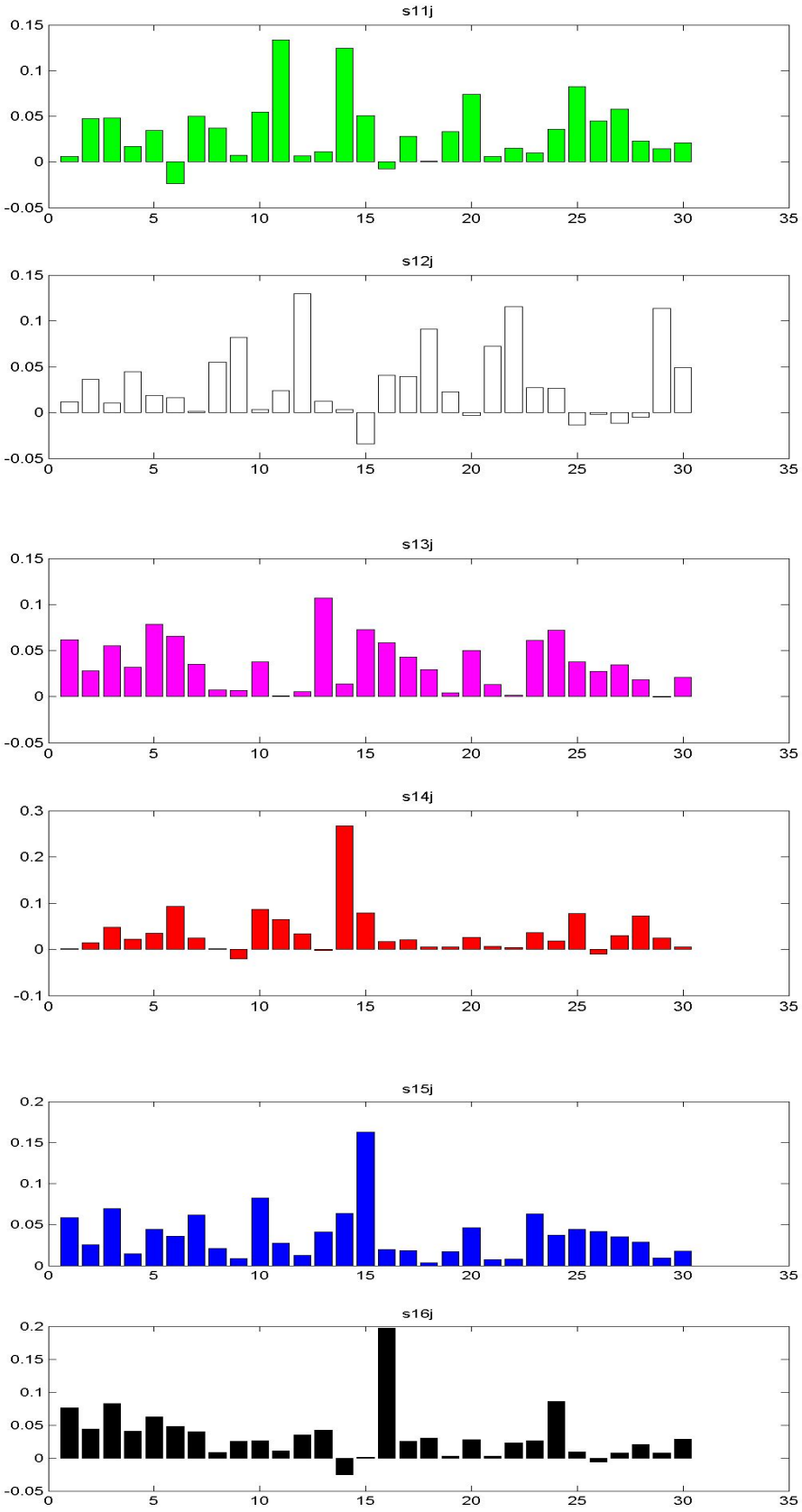


Figura 5.12.3 – Score dos locutores 11 ao 16 de 30, com 16 centróides.

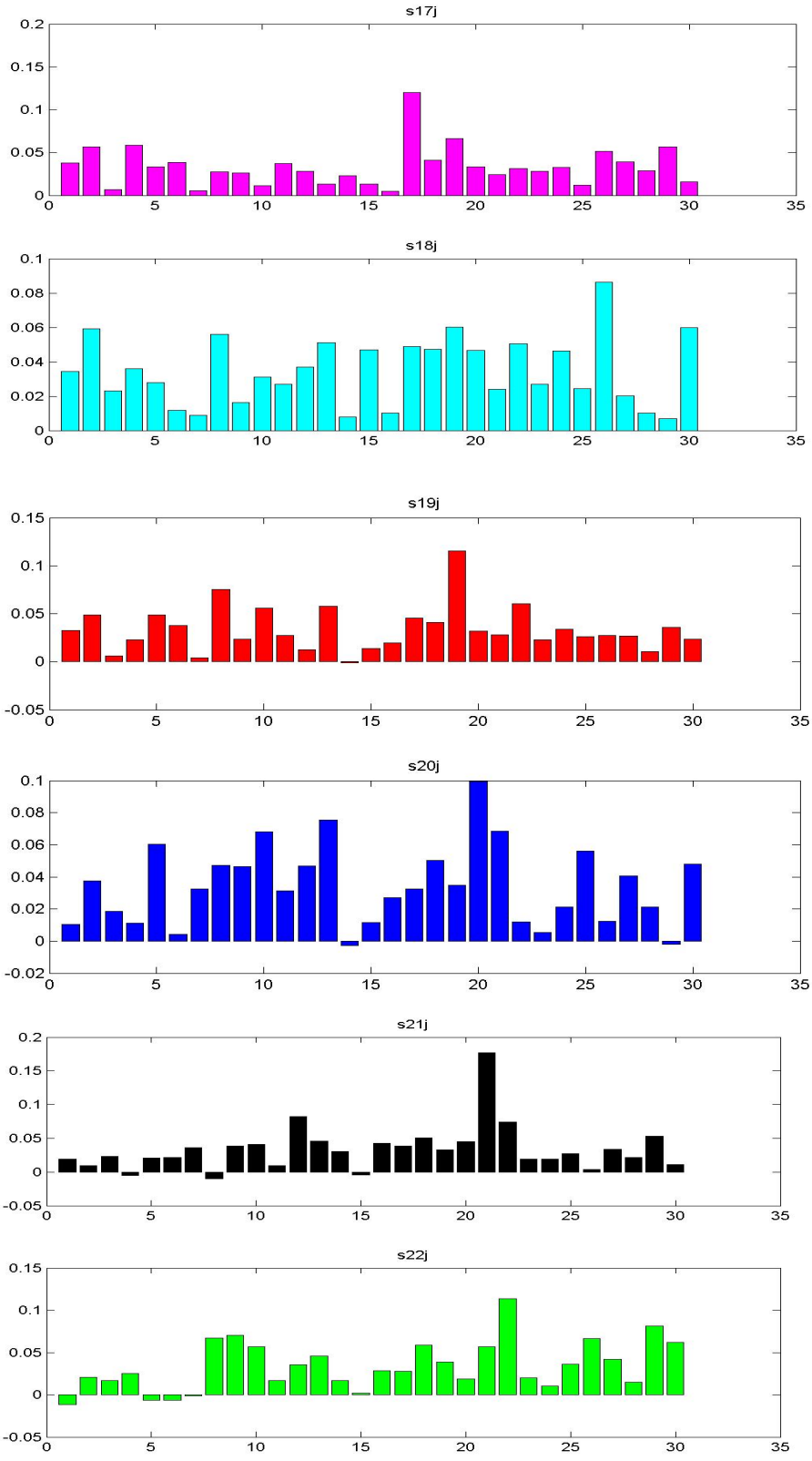


Figura 5.12.4 - Score dos locutores 17 ao 22 de 30, com 16 centróides.

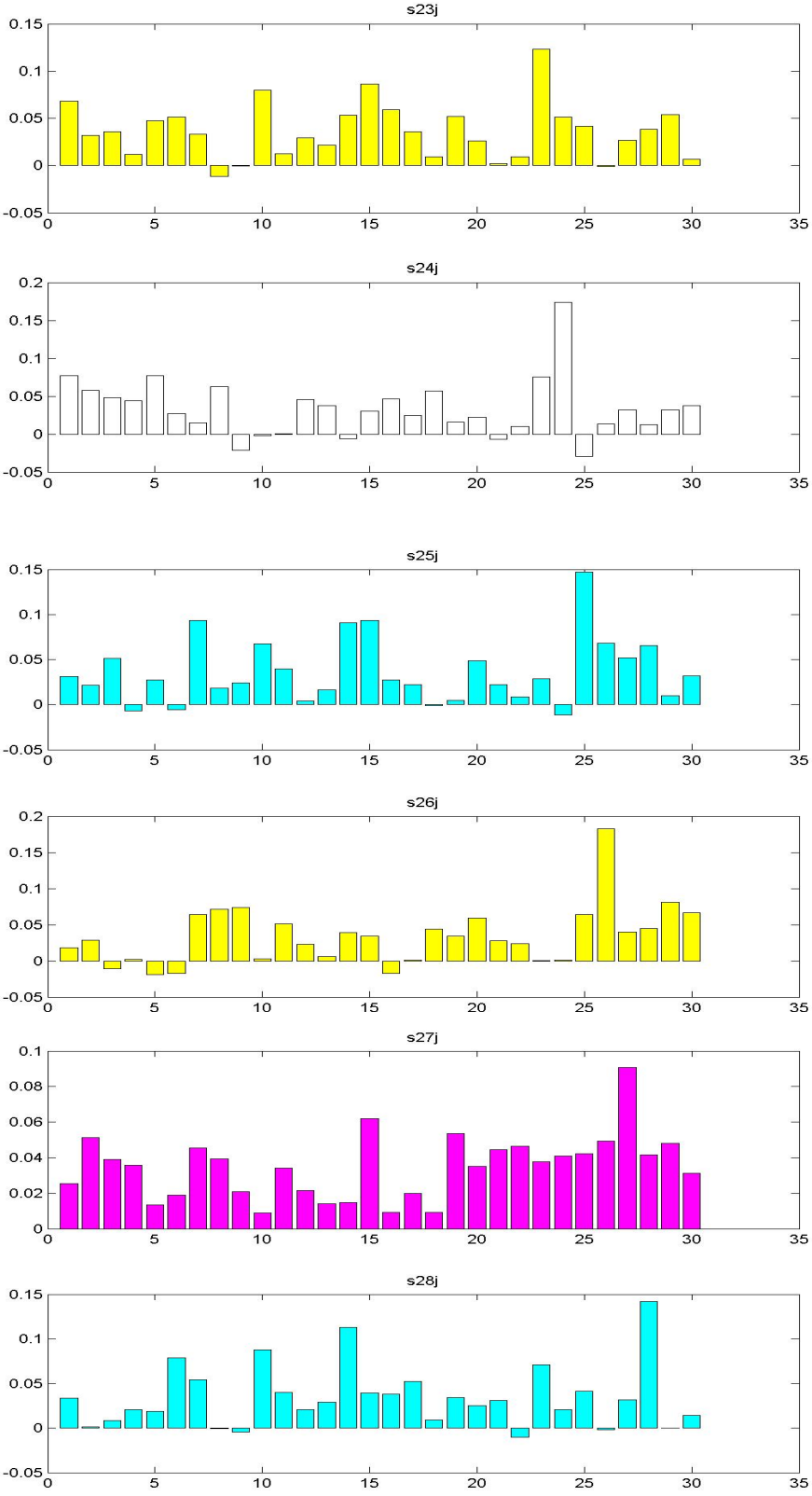


Figura 5.12.5 – Score dos locutores 27 e 28 de 30, com 16 centróides.

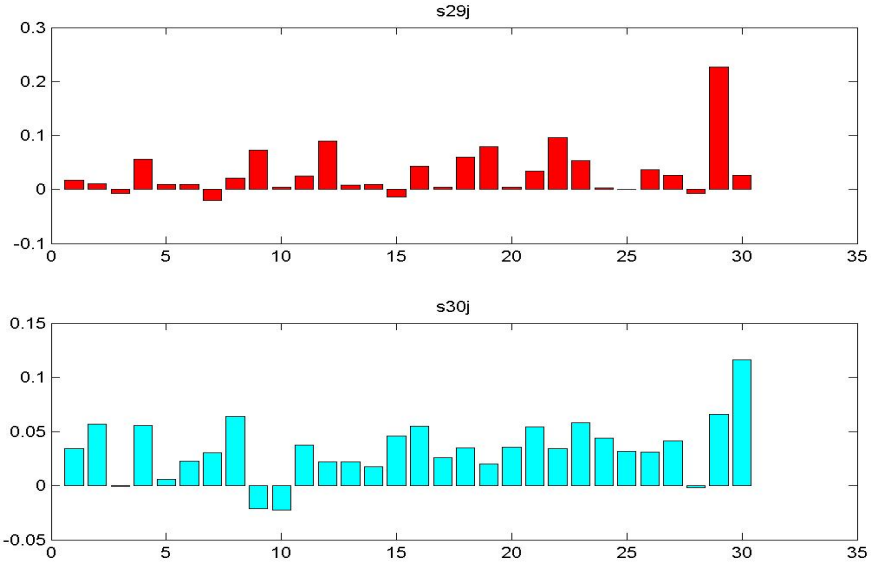


Figura 5.12.6 - Score dos locutores 29 e 30 de 30, com16 centróides.

Assim usando 16 centróides obtém-se o menor erro em 3,34%. Porém, agora é feito uma variação do número de centróides conservando-se o sistema com 30 locutores e verifica-se como ele se comporta quando os codebooks são calculados com 4 e 8 centróides**.** As Figuras 5.13 mostram o desempenho do sistema com codebooks de 4 centróides.

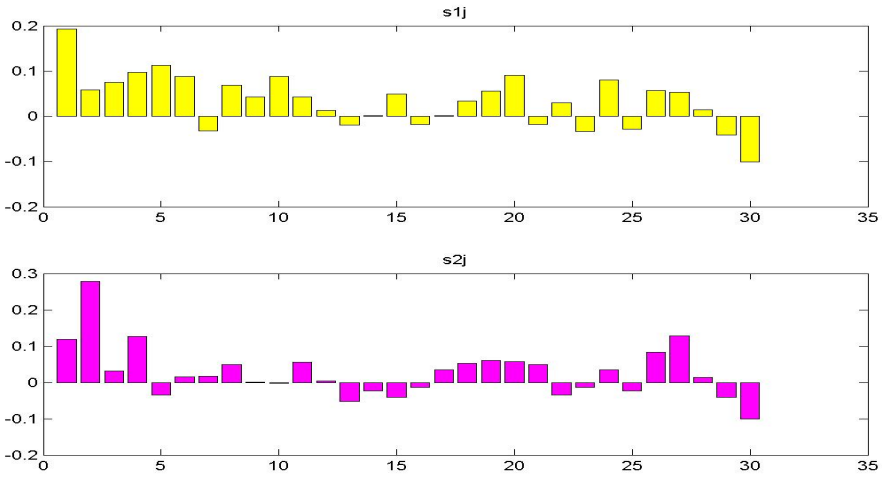


Figura 5.13.1 - Score dos locutores 1 e 2 de 30, com 4 centróides.

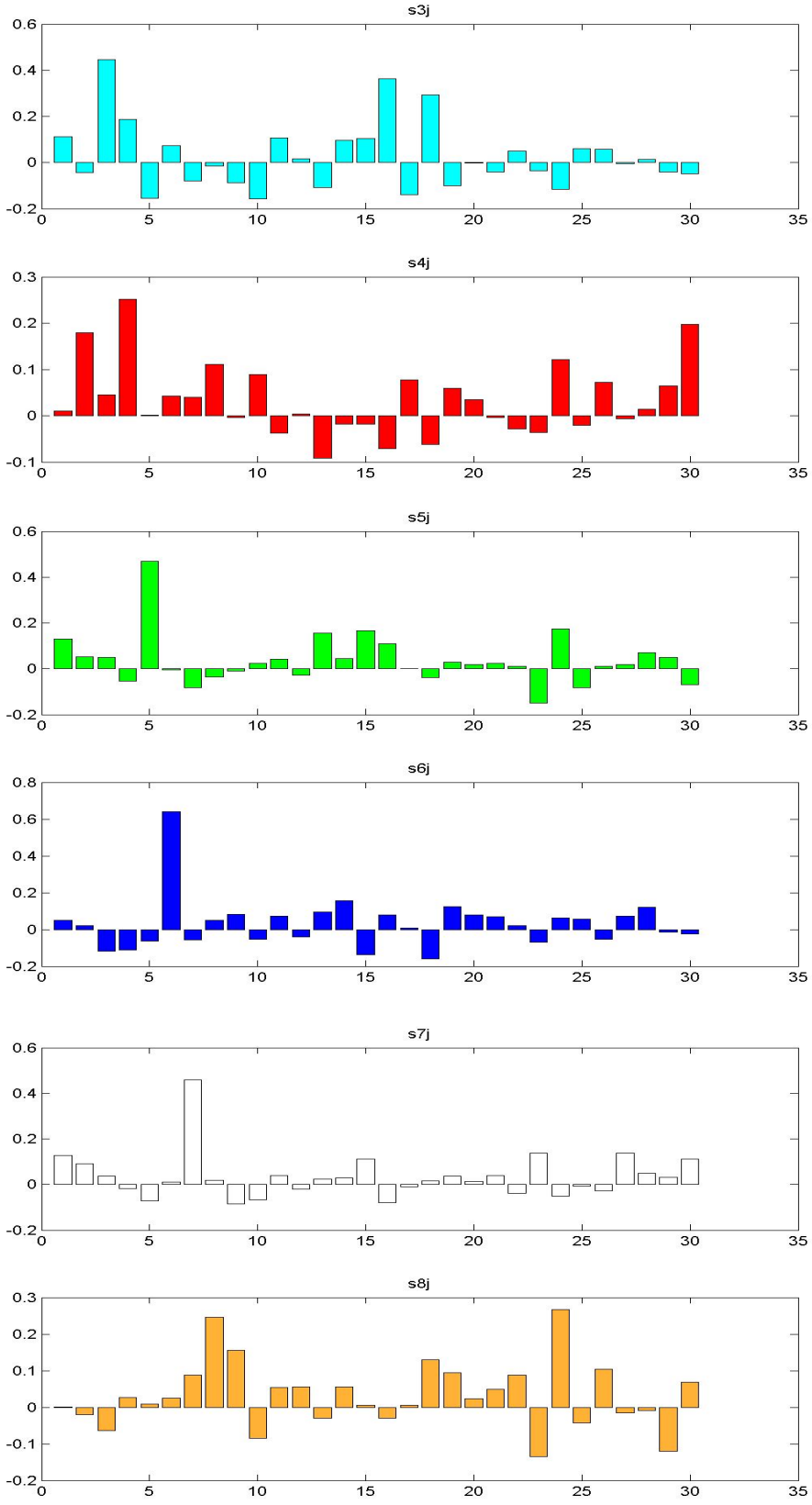


Figura 5.13.2 – Score dos locutores 3 ao 8 de 30, com 4 centróides.

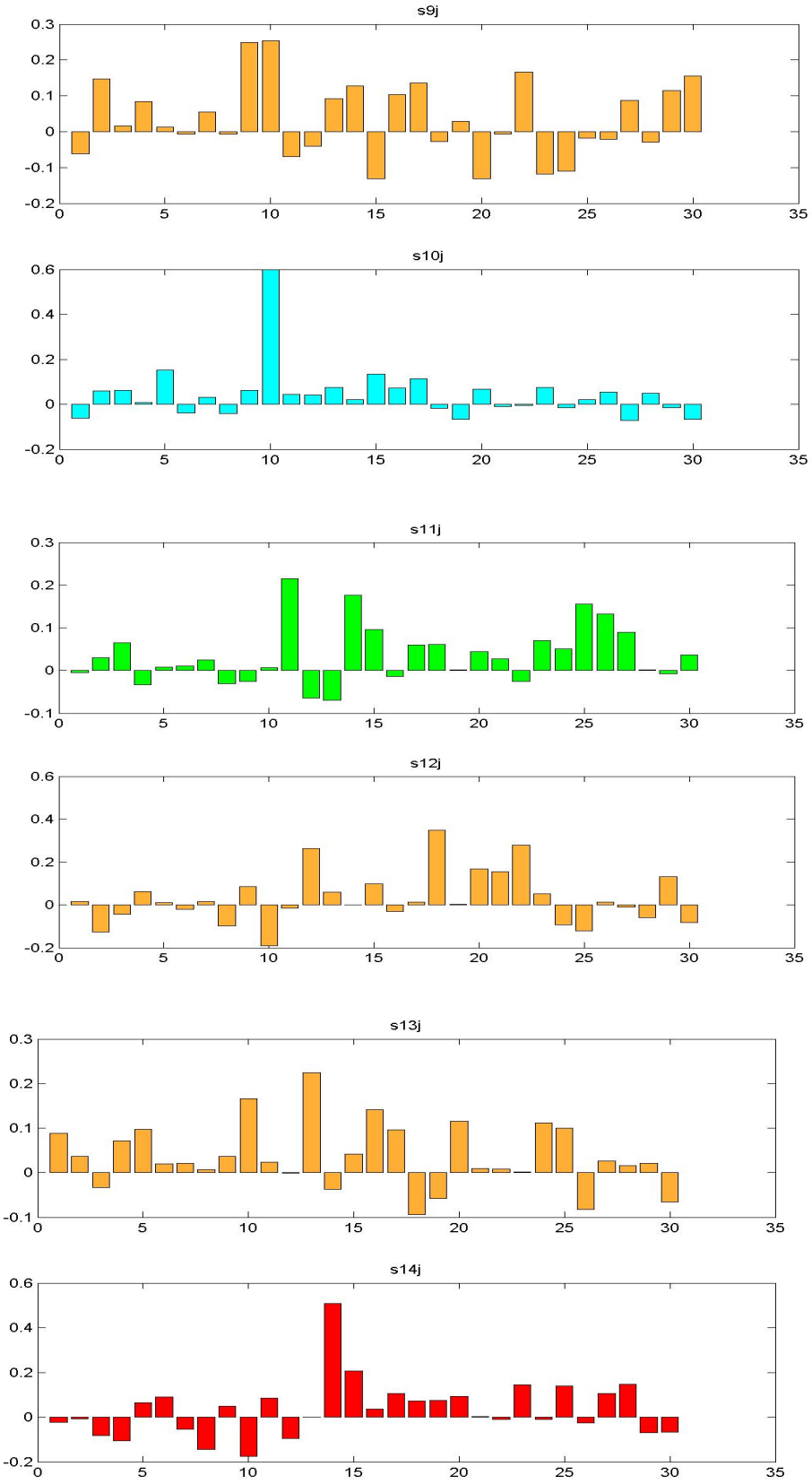


Figura 5.13.3 – Score dos locutores 9 ao 14 de 30, com 4 centróides.

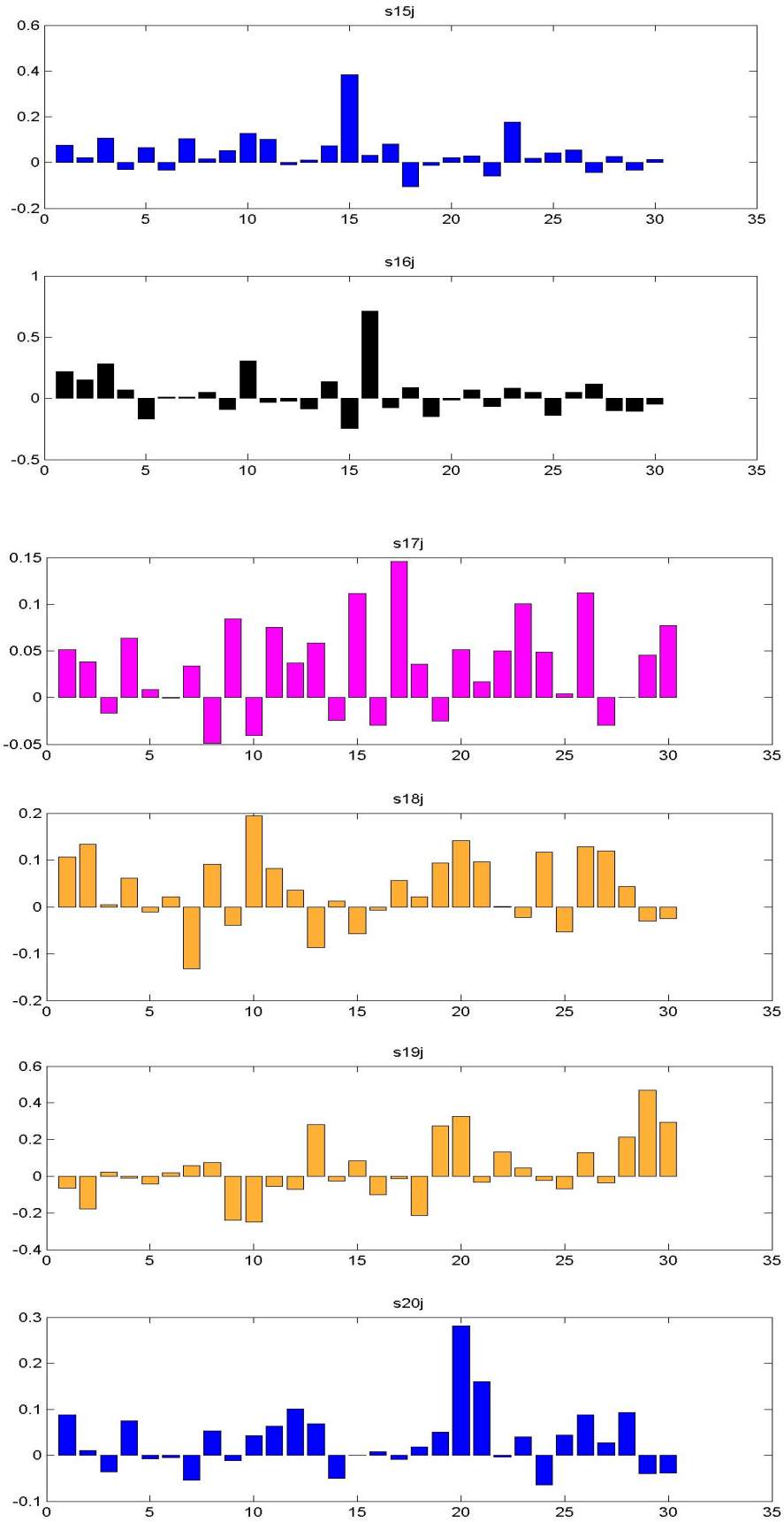


Figura 5.13.4 – Score dos locutores 15 e 20 de 30, com 4 centróides.

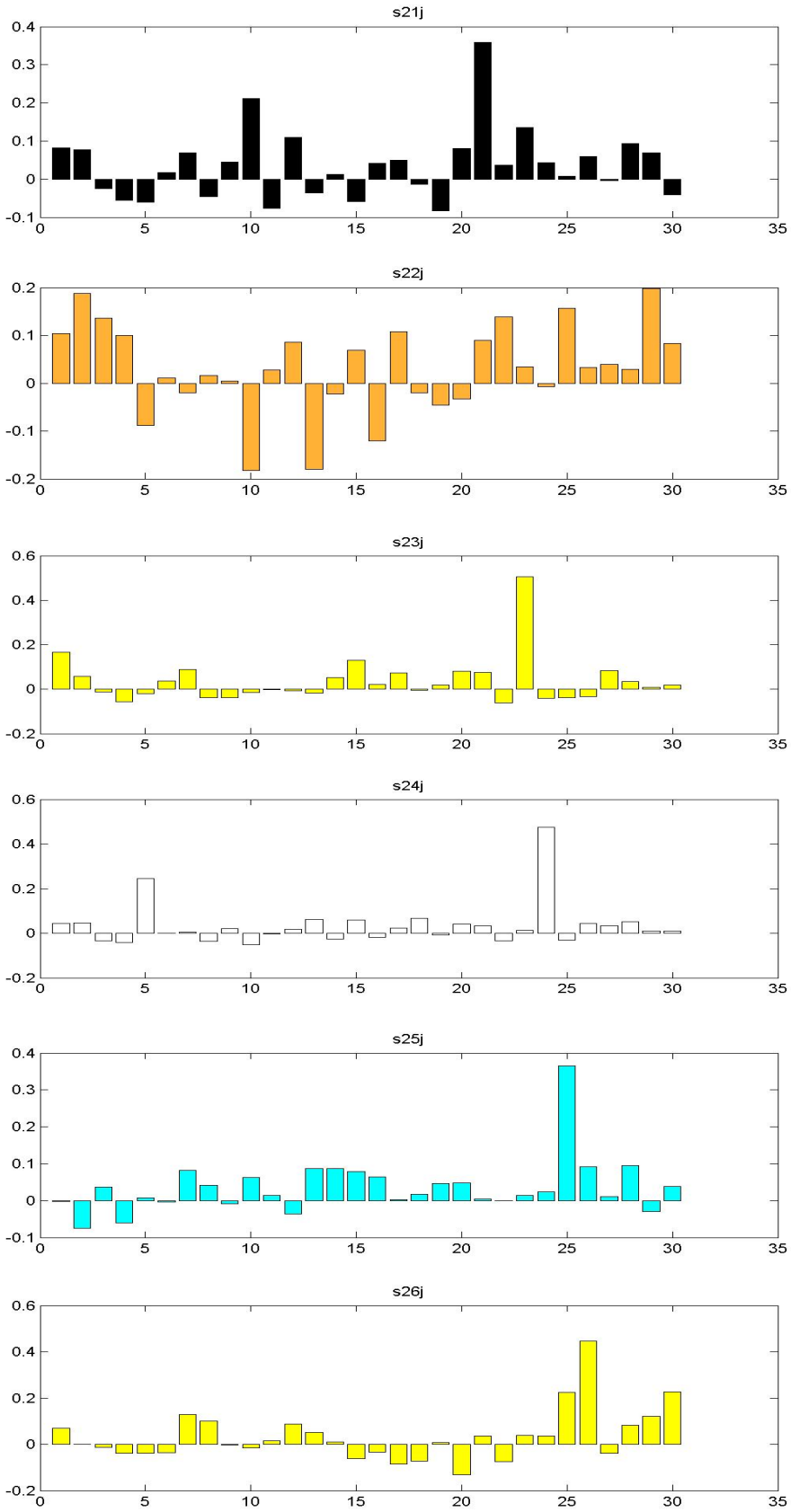


Figura 5.13.5 – Score dos locutores 21 ao 26 de 30, com 4 centróides.

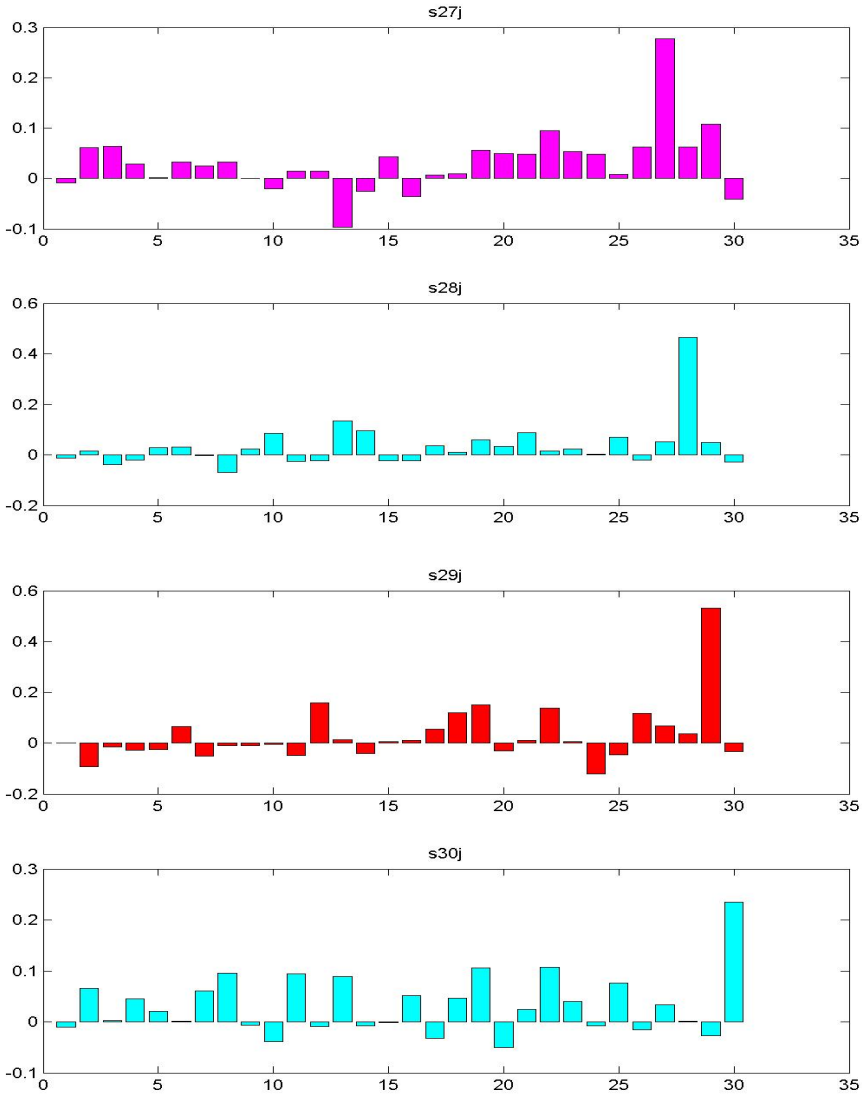


Figura 5.13.6 - Score dos locutores 27 ao 30 de 30, com 4 centróides.

Nota-se uma incidência de erro em torno de 23,34%, ocorridas em s_{ij}, para *i=8,9,12,13,18,19* e *22* respectivamente classificados como *j = 24,10,18,20,20,29* e *29*. Fazendo uma variação no número de centróides, de 4 para 8 e mantendo as outras estruturas ocorreu uma redução do erro para 13,34 %. Ocasionadas em $\boldsymbol{s_{ij}}$, com *i = 13,18,19* e *22* classificados respectivamente em *j = 23, 26, 10* e *26*.

O desempenho do sistema utilizando 8 centróides pode ser verificado nas Figuras 5.14 .

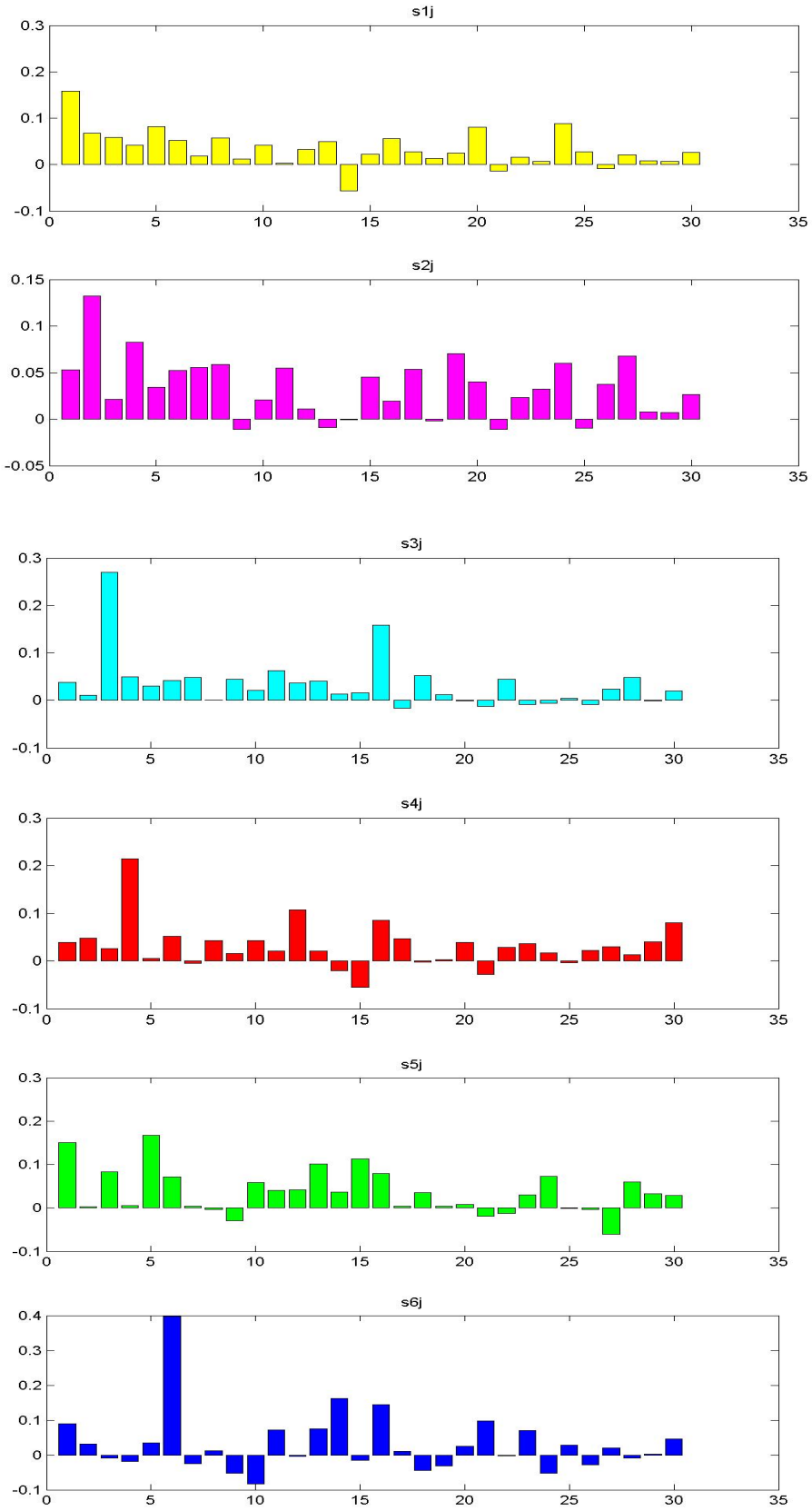


Figura 5.14.1 – Score dos locutores 1 ao 6 de 30, com 8 centróides.

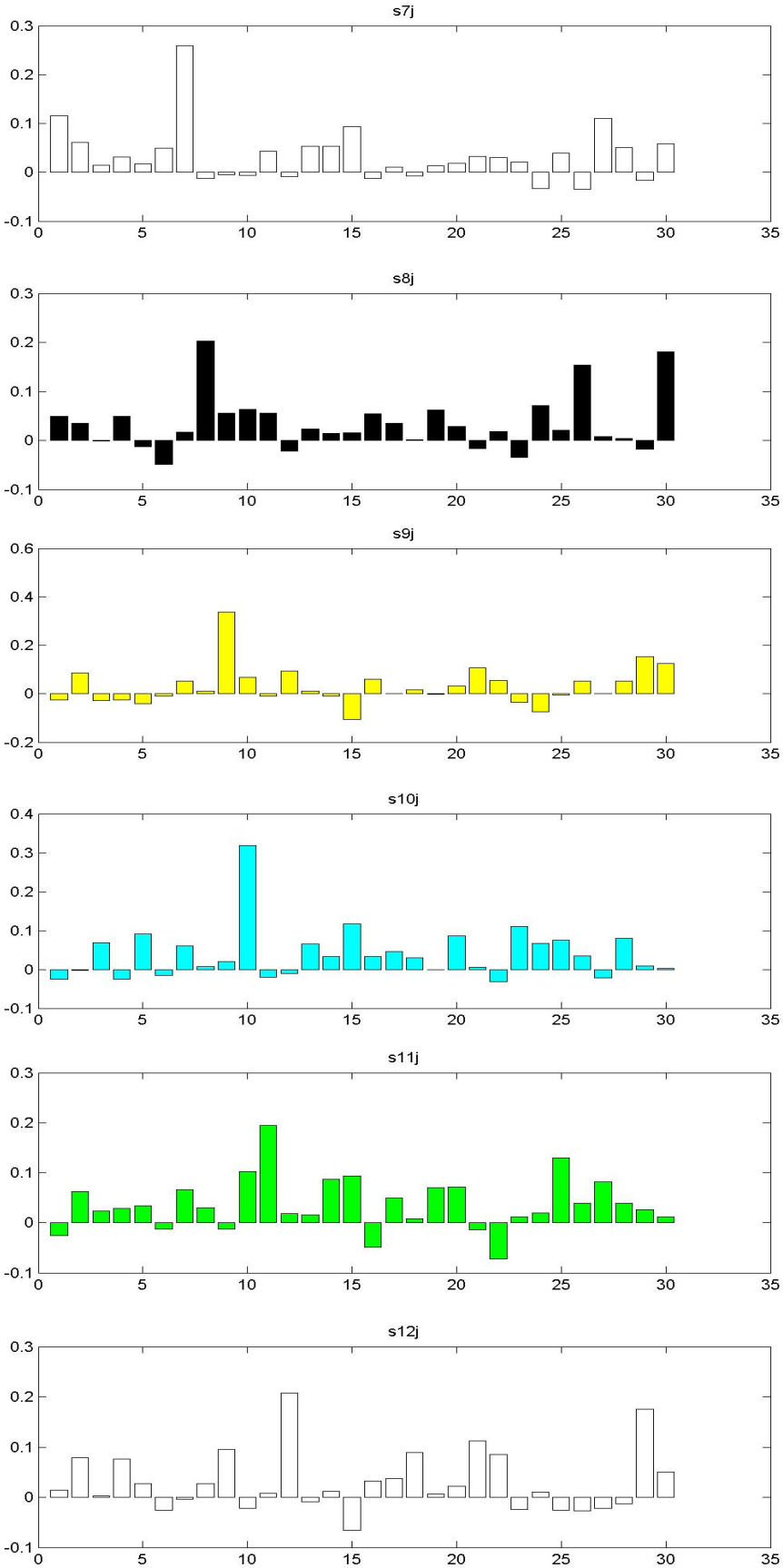


Figura 5.14.2 – Score dos locutores 7 ao 12 de 30, com 8 centróides.

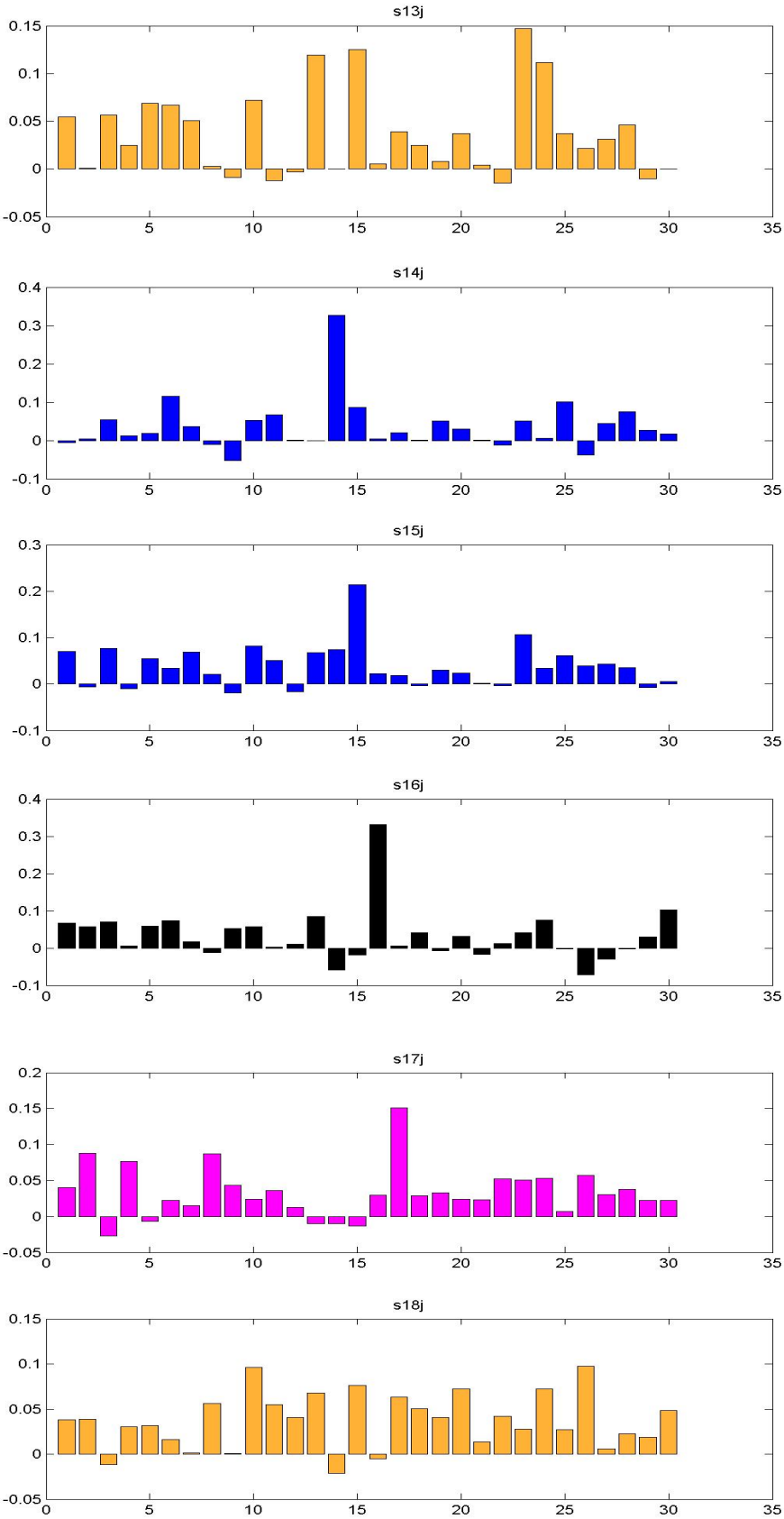


Figura 5.14.3 – Score dos locutores 13 ao 18 de 30, com 8 centróides.

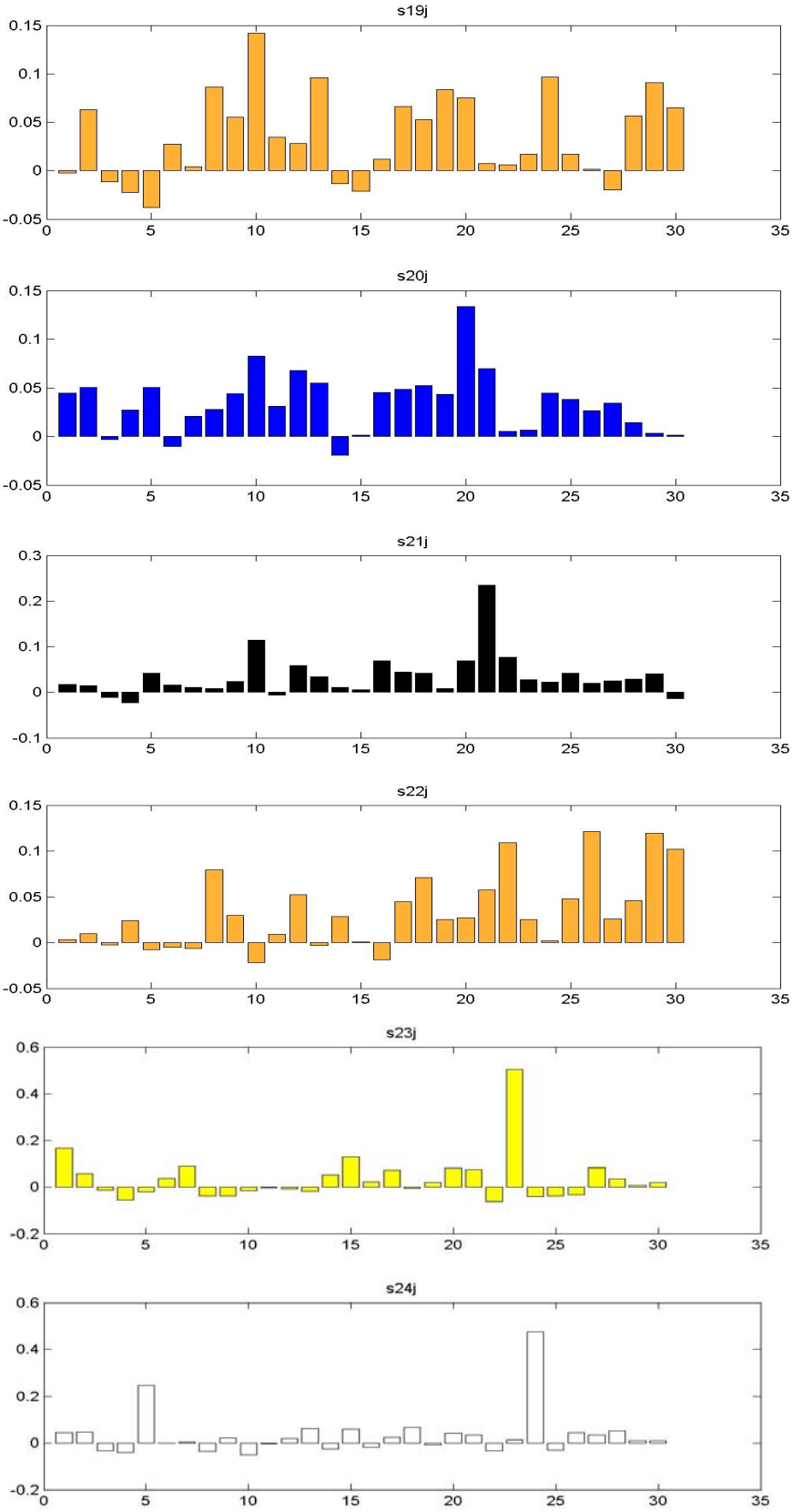


Figura 5.14.4 – Score dos locutores 19 ao 24 de 30, com 8 centróides.

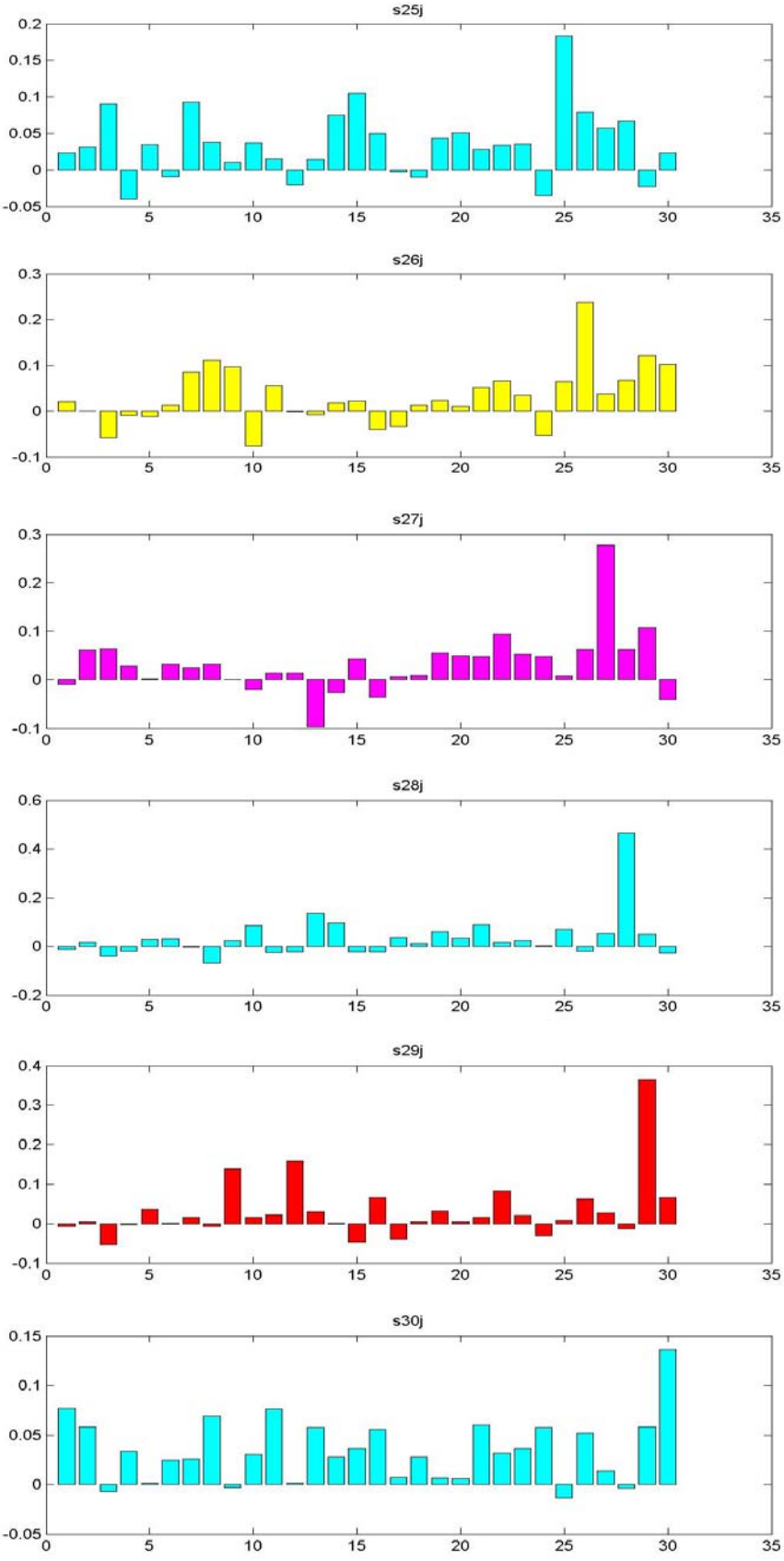


Figura 5.14.5 – Score dos locutores 19 e 30 de 30, com 8 centróides.

Assim pode-se resumir a ocorrência dos erros na Tabela 5.4. e Tabela 5.5.

TABELA 5.4 – Ocorrência do erro na identificação com a variação do número de centróides e locutores.

		NÚMERO DE LOCUTORES					
		5	**10**	**15**	**20**	**25**	**30**
CENTRÓIDES	**4**	***	***	***	***	***	**23,34%**
	8	***	***	***	***	***	**13,34%**
	16	**0%**	**0%**	**0%**	**5%**	**4%**	**3,34%**

(***) Configuração não testada.

Tabela 5.5- Locutores não identificados

		LOCUTORES NÃO IDENTIFICADOS
CONFIGURAÇÃO	**30 LOCUTORES E 4 CENTROIDES**	***8,9,12,13,18,19 e 22***
	30 LOCUTORES E 8 CENTROISDES	***13,18,19 e 22***
	30 LOCUTORES E 16 CENTROIDES	***18***
	25 LOCUTORES E 16 CENTROIDES	***18***
	20 LOCUTORES E 16 CENTROIDES	***18***

5.7 – Comparações entre alguns Métodos

Foi visto no capítulo I deste trabalho, uma breve noção a respeito dos vários métodos propostos para realizar a identificação e a verificação do locutor. Entre alguns desses métodos é feita uma comparação, mostrada na Tabela 5.6.

Tabela 5.6 - Comparação entre técnicas.

Proposto por	**Descrição do Projeto**	**Método**	**Erro**
Soong, et al. [27]	População: 100 Característica: LPC e CodeBook com 64 centróides Texto Dependente	QV	5%
Higgins and Wolford. [15]	População: 1 Característica: Ceptrum Texto Independente	DTW	10%
Reynolds [24]	População: 416 Característica: Mel-Cpetrais Texto Independente	HMM	11%
Figueiredo [11]	População: 10 Característica: LPC e CodeBook com 16 centróides Texto Dependente	QV	15%

Vejam que comparando os erros nas Tabelas. 5.4 e 5.6 pode-se verificar que o desempenho do método proposto neste trabalho é superior ao alcançado por [11] que possuía um configuração mais próxima e ao método proposto em [27] apesar do uso 64 centróides em seus codebooks. Quanto aos métodos propostos por [15] e [24] era esperado um erro inferior já que esses usam um texto independente.

5.8 – Considerações Finais deste Capítulo

De acordo com a Tabela 5.4 o processo de identificação não possui erro algum quando trabalha-se com *codebooks* com 16 centróides e um número pequeno de locutores, isto é, até 15 locutores e o erro é de 5% ao utilizar 20 locutores e cai a medida que se aumenta o número de locutores e sendo de 3,34 % quando se utiliza no processo os 30 locutores do banco de voz. O parece se tratar de uma incoerência, não é. De acordo com a configuração usada nos testes, alguns dos locutores não eram identificados.

Na Tabela 5.5 pode-se perceber a ocorrência de erro na identificação do locutor 18 para todas as configurações utilizadas. Isto permite saber que ocorreu um problema envolvendo as amostras deste locutor. Mas não foi feito testes suficientes para garantir este fato apenas 3 outros testes foram feitos usando 3 outros locutores não pertencentes ao banco de dados apresentado neste capítulo e o nível de acerto foi a 100% quando substituído pelo locutor 18, já que não foi possível re-gravar outros 3 arquivos deste, na configuração do sistema com 16 centróides e 30 locutores.

A implementação da verificação de locutores é notoriamente mais simples que a identificação. E a tabela de erros envolvendo a verificação dos locutores, Tabela 5.3, exibe o menor erro na eficiência do projeto com a utilização de classificadores polinomiais, principalmente quando o número de centróides é pequeno. Mas grande o problema da verificação está em estabelecer um limiar ótimo que melhor classificar os locutores, já que estão intimamente ligados aos *scores* obtidos pelos locutores aptos. A fórmula usada para calcular trata-se de uma sugestão deste trabalho que obteve bons resultados na sua implementação. Já que, a média de aptidão e o ponto médio entre as médias de aptos e inaptos não produziram os resultados desejáveis.

Na prática tem-se para os sistemas de verificação um limiar para cada um dos locutores com aptidão, e a comparação é feita de um pra um com resposta “sim” ou “não”.

No próximo capítulo serão levantadas as conclusões gerais a cerca deste trabalho.

CAPÍTULO VI

CONCLUSÕES, CONTRIBUIÇÕES E FUTUROS TRABALHOS

6.1 – Introdução

É apresentado neste capítulo uma conclusão geral, ou seja, uma análise global do trabalho. O objetivo deste trabalho era empregar classificadores polinomiais combinados à quantização vetorial para realizar a verificação e a identificação de um locutor.

Foi estudada a fisiologia e a anatomia da fala de forma a intensificar o fato já conhecido e existência de diferenças entre as vozes humanas, características que são muito importantes na realização do reconhecimento e na verificação do locutor. Um modelo físico matemático foi apresentado.

Posteriormente foi estudado a respeito do processamento digital de sinais, onde representou-se um sinal de voz, em termos de suas características, LPC. E foi aplicado os métodos quantificadores para reconhecer e identificar um grupo de locutores.

Para tal trabalho foi necessário o estudo das propriedades de classificação dos classificadores polinomiais e da quantização vetorial. Sua eficiência na separabilidade de um espaço *S*.

Um banco de dados formado de 30 pessoas e rotinas em *MatLab* foram desenvolvidas para que os testes comprobatórios fossem obtidos. Estes realizados no domínio do tempo, em que se obteve um nível de verificação e de reconhecimento em torno de 3,34 % em ambos os casos.

6.2 – Contribuições deste Trabalho

Neste trabalho desenvolveu-se um estudo a respeito da combinação de dois métodos, quantização vetorial e classificadores polinomiais, aplicados à verificação e à identificação de locutor pela sua voz. Cuja técnica obteve-se bons resultados.

Desenvolvimento de um Banco de Dados formados por *coodeboks* usando uma variedade de centróides.

Desenvolvimento de funções usando *MatLab* que poderão auxiliar em trabalhos futuros envolvendo análise polinomial.

6.3 – Futuros Trabalhos

O presente trabalho permite a continuidade de pesquisas em direções ainda não exploradas, tais como:

1. Aplicabilidade deste método para um banco de dados maior que o apresentado

neste trabalho.

2. Exploração de polinômios com graus superiores, onde é fato que trará resultados ainda melhores que os apresentados neste trabalho.
3. Comparação entre as diferentes expansões polinomiais de mesmo grau, trabalho que envolve muita intensidade matemática para formalização dos resultados.
4. Uso de outras características tais como coeficientes Mel-Ceptral e Cepstral.
5. Uso de transformadas *Wavelets* combinado aos classificadores polinomiais.
6. Utilização de outro tipo de janelamento.
7. Implementação usando *C* em *DSP* para desenvolvimento de um hardware para análise em tempo real.
8. Medida da eficiência do método utilizando texto independente.
9. Viabilidade quanto à aplicação em imagens.

6.4 - Considerações Finais

Muito deve ser explorado a respeito da aplicabilidade de classificadores polinomiais, graus, natureza e sua eficiência quanto a separabilidade. Não apenas envolvendo identificação e verificação de locutores pela sua voz, mas outros padrões porque o ser humano a todo o momento classifica objetos mesmo que inconscientemente e estabelece estruturas organizacionais de classificação. Cada pesquisa, e cada trabalho científico é apenas o *"start"* para os caminhos a serem trilhados durante a infinita evolução tecnológica.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ASSALEH K. T. et CAMPBELL W. M., Speaker Identification Using a Polynomial Based Classifier, FIFTH INTERNATIONAL SYMPOSIUM ON SIGNAL PROCESSING AND ITS APPLICATIONS, Brisbane, Australia, p.115-118, 22-25, Aug. 1999.

[2] ATAL B.S., Automatic Recognition of Speakers from their Voices, Proceedings of the IEEE, v.64, n.4, p. 460-474, Apr. 1976.

[3] BARNEY, A., SHANDLE, C. H., and DAVIS, P. O. A. L., Fluid Flow in a Dynamical Mechanical Model of the Vocal Folds Tract. 1: Measurements and Theory, J. Acoustical Society of America, v. 105, n.1, p. 444-455, Jan. 1999.

[4] CAMPBELL, J. P. Jr. Speaker Recognition: A Tutorial. Proceedings of the IEEE, v.85, n. 9, p.205-212, Sep.1997.

[5] CAMPBELL J. P. Jr. and REYNOLDS, D. A., Corpora for the Evaluation of Speaker Recognition Systems. Proceedings of the IEEE, p.829-832, Mar. 1999.

[6] CAMPBELL, W.M. and ASSALEH, K.T. Polynomial Classifier Techniques for Speaker Verification. Proceedings of the IEEE, p.321-324, May 1999.

[7] CAMPBELL, W.M. and ASSALEH, K.T. Speaker Recognition with Polynomial Classifiers. Proceedings of the IEEE, v.10, n.. 4, p.205-212, May 2002.

[8] COSTA FILHO, A.C. Reconhecimento e Segmentação de Fonemas na Fala Contínua Utilizando Quantização Vetorial. Uberlândia: Universidade Federal de Uberlândia, Faculdade de Engenharia Elétrica, p. 24-39, 1996.

[9] DELLER, J.R.; POAKIS, J.G., and HANSEN, J.H. Discrete – Time Processing of Speech Signals. New York: Macmillan,. p.908 ,1993

[10] DENES, P. B. and PINSON, E. N., The Speech Chain: The Physics and Biology of Spoken Language, Anchor Press-Doubleday, Garden City, NY, 1973.

[11] FIGUEIREDO, S. A. Contribuição ao Estudo do Reconhecimento do Locutor Pela Voz. Uberlândia: Universidade Federal de Uberlândia, Faculdade de Engenharia Elétrica, 1990.

[12] FLANAGAN, J. L., Speech Analysis, Synthesis, and Perception, Second Edition, Spring-Verlag, New York, NY, 1972.

[13] FLANAGAN, J. L. and ISHIZAKATA, K., Computer Model to Characterize the Air Volume Displaced by the Vibrating Vocal Cords, J. Acoustical Society of America, v. 63, n. 5, p.1559-1565, May 1978.

[14] GRAY, R. M., BUZO, A., GRAY JR., A.H. and MATSUYAMA, Y., Distorsion Measures for Speech Processing , IEEE, Transaction on Acoustics, Speech and Signal Processing, v. ASSP-28, p.367-376, Aug. 1980.

[15] HIGGINS A., BHALER L., and PORTER, A new method of text- Acoustic, Speech, and Signal Processing, Tokyo, Japan, 1996, p.869-872.

[16] KRANE, M., SINDER, D., and FLANAGAN J., Synthesis of Unvoiced Speech Sounds Using an Aero acoustic Source Model, J. Acoustical Society of America, v. 105, n. 2, pt. 2, ASA Convention Record, p. 1160, Feb. 1999.

[17] LINDE, Y., BUZO, A. e GRAY, R.M. Na Algorithm for Vector Quantizer Design. IEEE Transactions on Communications, v. COM –28, n. 1, p.84-95, Jan. 1980.

[18] MAKHOUL, J., Linear Prediction: A Tutorial Review, Proceedings of IEEE, v. 63, n. 4, p. 561-580, Apr. 1975.

[19] MAKHOUL, J.; ROUCOS, S. and GISH, H., Vector Quantization in Speech Coding, Proceedings of the IEE, v. 73, n. 11, p 1551-1558, Nov. 1980.

[20] PARSONS, T. W., Voice and Speech Processing, New York, McGraw Hill, p.402

[21] PEACOCKE, R. D. et GRAF, D. H., An Introduction to Speech and Speaker Recognition, Computer, p. 26-33 , Aug. 1990.

[22] QUATIERI, T. E., Discrete- Time Speech Signal Processing. Prentice Hall. Hancover, Oct 2001, p.55-72.

[23] RABINER, L. R. and SCHAFER R.W., Digital Processing of Speech Signals, New Jersey: Prince Hall, Inc., 1978. p.512.

[24] REYNOLDS, D. MIT Lincoln Laboratory site presentation, in Speaker Recognition Workshop, A. Martin, Ed. Sect.5, Maritime Of Technology, Linthicum Heights, MD, Mar.27-28, 1996.

[25] ROSENBERG, A. E., Automatic Speaker Verification: A Review, Proceedings of the IEEE, v. 64, n. 4, p. 475-487, Apr. 1976.

[26] SAKOE, H. et CHIBA, S, Dynamic Programming Algorithm for Optimization for Spoken Word Recognition, IEEE Trans. ASSP, ASSP – 26, p. 621-659, 1981.

[27] SONG, F. K., ROSEMBERG, A.; HUANG, B. H. and RABINER, L. R., A Vector Quantization Approach to Speaker Recognition, AT & T Technical Journal, v. 66, ISSUE 2, p. 14-26, Mar/ Apr 1987.

[28] STEVENS, K. N., Acoustic Phonetics, The MIT Press, Cambridge, MA, 1998.

[29] STORY, B. and TITZE, I. R., Voice Simulation with a Body Cover Model of the Vocal Folds, J. Acoustical Society of America, v. 97, p. 1249-1260, 1995.

[30] TEAGER, H. M. and TEAGER S. M., Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract, chapter in Speech Production and Speech Modeling, W. J. Hardcastle and A. Marchal, eds., NATO Adv. Study Inst. Series D, v.55, Bonas France; Kluwer Academic Publishers, p. 241-261, Boston, MA, 1990.

[31] WAN, V., RENALS S. Evaluation of Kernel Methods for Speaker Verification and identification. Proceedings of the IEEE, p.669-672, 2002.

BIBLIOGRAFIA

[1] DUDA, R. O. and HART P.E., Pattern Cassification and Scene Analysis. Stanford Research Institute, Menlo Park, California, John Wiley & Sons, 1973.

[2] GOLUB, G.H. and VAN LOAN, C.F., Matrix and Computations, Baltimore: Jonh Hopkins, 1989.

[3] JANG, G., LEE, T. and OH, Y., Learning Statistically Efficient Features for Speaker Recognition, Proceedings of the IEEE, p. 437-440, Apr. 2001.

[4] LATHI, B.P., Linear Sytems and Signals, New York, Oxford Univerty Press, 2005.

[5] LUMMIS, R.C., Speaker Verification by Computer Using Speech Intensity for Temporal Registration. Proceedings of the IEEE, v.AU-21, n.. 2, p. 80-88, Apr 1973.

[6] NAIK, J.M., Speaker Verification: A Tutorial. Proceedings of the IEEE, p.42-48, Jan. 1990.

[7] OPPENHEIN, A.. Reconhecimento e Segmentação de Fonemas na Fala Contínua Utilizando Quantização Vetorial. Uberlândia: Universidade Federal de Uberlândia, Faculdade de Engenharia Elétrica, p. 24-39. 1996

[8] ZHANG, X, WU, J. and ZHANG, Q., Speaker Identification Based Modified Polynomial Classifiers. Proceedings of the IEEE, p.3178-3182, Mar 1999.

ANEXO

PROGRAMAS E FUNÇÕES DO MATLAB

As Simulações foram executadas via programas desenvolvidos para o MatLab, devido a grande facilidade e a potencialidade na manipulação de vetores e matrizes. Todos os programas pode ser solicitados ao autor pelo email: *wdparreira@yahoo.com.br*.

Segue a Relação de Programas utilizados:

Arquivo (*.m)*	**Função**	**Observação**
Qvet	Calcula os Codebooks dos arquivos *.wav*	*função*
getcbook	Cria arquivos do banco de dados formado pelos codebooks gerados por *Qvet.m*	*função*
signal2frames	Lê os arquivos *.wav*, faz divisão em frames a aplica o *Janelamento de Hamming*	*função*
classipoly5	Faz a identificação usando 5 locutores do banco de dados.	
classipoly10	Faz a identificação usando 10 locutores do banco de dados.	
classipoly15	Faz a identificação usando 15 locutores do banco de dados.	
classipoly20	Faz a identificação usando 20 locutores do banco de dados.	
classipoly25	Faz a identificação usando 25 locutores do banco de dados.	
classipoly30	Faz a identificação usando 30 locutores do banco de dados	
classipolyverif	Faz a verificação dos Locutores	

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo