

# CARACTERIZAÇÃO ACÚSTICA DAS VOGAIS DO PORTUGUÊS BRASILEIRO VISANDO A NORMALIZAÇÃO DE LOCUTORES

Gustavo F. Rodrigues      Hani C. Yehia

gustavo@cpdee.ufmg.br      hani@cefala.org

CEFALA - Center for Research on Speech, Acoustics, Language and Music

UFMG - Graduate Program on Electrical Engineering

Av. Antônio Carlos, 6627, Campus - Pampulha

CEP 31270-901 Belo Horizonte-MG - Brasil

**Resumo** – No desenvolvimento de sistemas reconhecadores de voz a alteração do trato vocal para diferentes locutores representa uma das principais fontes de variações entre locutores diversos imposta ao sinal da fala. Desde então, várias técnicas de normalização de locutor vem sendo propostas com o objetivo de reduzir a variabilidade entre locutores. O propósito deste artigo consiste em estudar as características e variações espectrais das vogais do português brasileiro (/a/,/e/,/i/,/o/,/u/,/ê/,/ô/) pronunciadas por diferentes locutores. Foi analisado a variância existente entre os diversos coeficientes: Linear Prediction Code (LPC), Partial Correlation (PARCOR), Line Spectrum Pair (LSP) e Mel-Frequency Cepstral Coefficients (MFCC), sendo estes parâmetros utilizados para a extração de dados de um sinal acústico. Esta análise visa identificar as técnicas de extração de dados que apresentam maior robustez quanto a variabilidade do locutor bem como a necessidade de normalização destes parâmetros.

**Abstract** – The differences in the vocal tract dimensions are the main responsible for the variability found in the speech signal. Many normalization methods have been proposed in order to reduce the variance among speakers. The purpose of this paper is to discuss the spectrum variation found in the oral vowels of Brazilian Portuguese (/a/,/e/,/i/,/o/,/u/,/ê/,/ô/). The variance present in among several parameters: Linear Prediction Code (LPC), Partial Correlation (PARCOR), Line Spectrum Pair (LSP) e Mel-Frequency Cepstral Coefficients (MFCC), which had been used to extract the data from the speech signal, is investigated. This study was done in order to identify the methods of data extraction that performe better to represent the variability of the vowels when they are spoken by different speakers.

## I. INTRODUÇÃO

Um reconhecedor de voz requer algum tipo de normalização antes de aceitar um novo locutor devido a variabilidade existente entre locutores. Sendo assim, um reconhecedor eficiente precisará identificar a variabilidade existente quando locutores diversos pronunciam por exemplo, as vogais, e desta forma promover a normalização das mesmas. Vários estudos [1] demonstram a importância dos dois ou três primeiros formantes na determinação da qualidade da vogal em relação as demais formantes. Sendo necessário somente as duas primeiras formantes para classificar as vogais dentro de um espaço bidimensional onde todas as vogais apresentam posições padrões que são similares para todos os locutores. Devido a diferença

entre as dimensões do trato vocal, dois locutores podem pronunciar vogais cujo sons são similares porém apresentam valores de formantes diferentes, assim como também podemos obter sons diferentes com valores de formantes similares [2]. A normalização consiste então em uma transformação (linear ou não-linear) dos dados do sinais acústicos para um conjunto de sons ou vogais pronunciadas por diferentes locutores. Isto com o objetivo de eliminar as características particulares de cada locutor do sinal fonético deixando somente a qualidade fonética comum a todos os locutores de uma determinada língua ou dialeto [3]. Existem atualmente diversos algoritmos para normalização [7] e [8]. No caso da normalização das vogais, o critério para estabelecimento do sucesso destes algoritmos se baseia na máxima redução da variabilidade entre cada grupo da mesma vogal e na manutenção ou aumento da separação entre os diversos grupos de vogais. Para extração das características e comportamento das vogais do português brasileiro foram obtidas gravações junto a um conjunto de 4 homens e 4 mulheres. As gravações realizadas apresentam as seguintes características:

- Foram gravadas com taxa de amostragem de 16.000 Hz.
- Os dados foram divididos em quatro grupos, sendo que no primeiro grupo estão as vogais pronunciadas isoladamente (/a/,/e/,/i/,/o/,/u/,/ê/ e /ô/) e nos demais grupos as vogais foram pronunciadas dentro de palavras contendo a estrutura consoante-vogal-consoante. Respectivamente as palavras que foram pronunciadas são as seguintes: ab-vogal-ba, ad-vogal-da, ag-vogal-ga.
- Os locutores pronunciaram três vezes cada vogal ou palavra.

## II. MEDIÇÃO FORMANTES

Na Figura 1 temos a representação da primeira formante versus a segunda formante para as vogais pronunciadas pelos seis locutores mencionados anteriormente.

Observe que na Figura 1 as regiões das vogais foram traçadas arbitrariamente visando englobar 90% de cada classe de vogal. A frequência de cada formante foi estimada junto ao *spectrogram* do sinal acústico através da média de sua frequência durante o intervalo de tempo de cada vogal. Para a extração dos formantes utilizou-se o software Praat desenvolvido por Paul Boersma [9]. Algumas dificuldades foram encontradas na fase de extração dos formantes, possivelmente devido a baixa relação sinal ruído (Signal Noise Ratio - SNR) das gravações que variaram entre 25 a 30 decibéis. Outras dificuldades encontradas foram os casos onde a frequência fundamental é alta provocando baixa definição das frequência dos formantes, conforme relatado em [2]. Na Figura 2 estão representadas as médias do segundo formante ( $F_2$ ) versus o primeiro formante ( $F_1$ ) para todas as vogais, sendo as mesmas separadas em dois grupos: masculino e feminino.

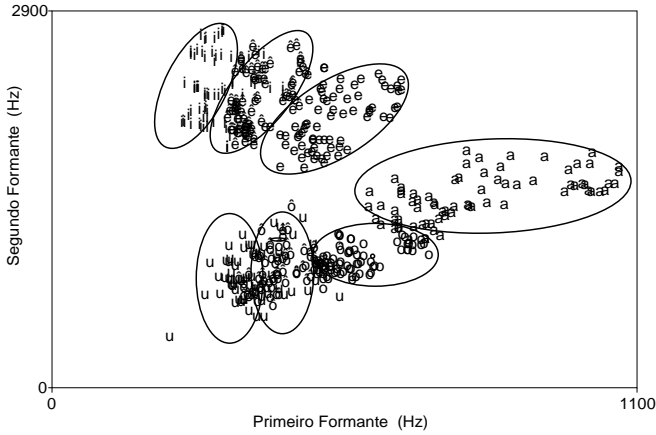


Figura 1. Frequência do segundo formante versus frequência do primeiro formante para as sete vogais pronunciadas por 6 locutores.

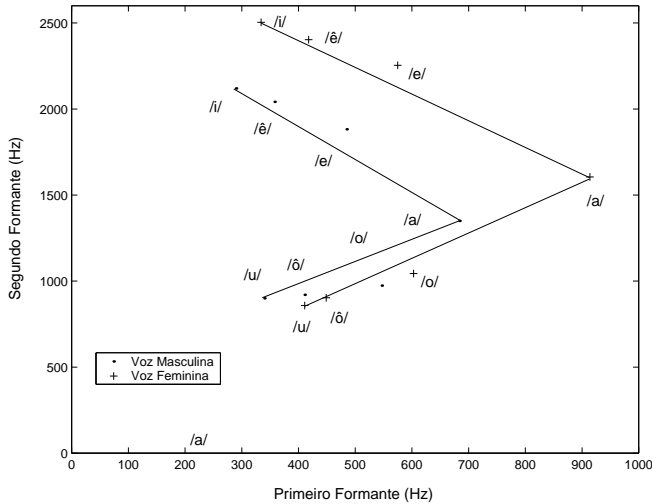


Figura 2. Comparação entre as médias de  $F_1$  e  $F_2$  obtidas entre os locutores homens e mulheres.

Uma análise mais detalhada dos dados mencionados pode ser encontrada na Tabela 1 que identifica a média ( $\bar{X}$ ), desvio padrão ( $\sigma$ ) e o coeficiente de variação ( $\nu$ ) entre os locutores para as três primeiras formantes ( $F_1$ ,  $F_2$  e  $F_3$ ).

É interessante notar neste ponto que de acordo com as figuras 1 e 2, as frequências dos formantes variam para um mesmo locutor e entre locutores diferentes. Porém tendem a ocupar uma região padronizada. Outro item relevante refere-se ao fato de observarmos através da Tabela 1 que os coeficientes de variação do terceiro formante  $F_3$  tendem a sofrer menor variação ao longo do tempo que os demais formantes  $F_1$  e  $F_2$ .

### III. EXTRAÇÃO DE PARÂMETROS

Em reconhecimento de fala, a fase de pré-processamento e extração das características de um sinal acústico é considerada uma etapa fundamental para o bom desempenho do sistema. Neste estágio a informação que é descartada é perdida para sempre ao mesmo tempo que um ruído que é aceito poderá afetar a performance do sistema de reconhecimento. Sendo assim, deve-se estar atento na escolha dos parâmetros que irão parametrizar o sinal de forma que estes parâmetros preservem ao máximo os aspectos relevantes ao sinal da fala e

TABELA I  
MÉDIA DAS FREQUÊNCIAS (HZ) DOS FORMANTES DAS VOGAIS]

Vogais		/a/	/e/	/i/	/o/	/u/	/ê/	/ô/
Homem								
$F_1$	$\bar{X}$	684	485	290	547	340	358	411
	$\sigma$	46	37	28	34	35	15	23
	$\nu$	0,07	0,08	0,10	0,06	0,10	0,04	0,06
Mulher								
$F_1$	$\bar{X}$	914	575	334	603	411	418	449
	$\sigma$	106	56	47	71	57	42	56
	$\nu$	0,08	0,10	0,14	0,12	0,14	0,10	0,12
Homem								
$F_2$	$\bar{X}$	1350	1882	2120	974	900	2042	920
	$\sigma$	121	97	139	99	179	95	173
	$\nu$	0,09	0,05	0,07	0,10	0,20	0,05	0,19
Mulher								
$F_2$	$\bar{X}$	1605	2254	2504	1044	858	2403	903
	$\sigma$	122	113	166	95	200	132	125
	$\nu$	0,06	0,05	0,07	0,09	0,23	0,05	0,14
Homem								
$F_3$	$\bar{X}$	2405	2598	2863	2713	2665	2689	2761
	$\sigma$	271	142	153	369	213	124	237
	$\nu$	0,11	0,05	0,05	0,14	0,08	0,05	0,09
Mulher								
$F_3$	$\bar{X}$	2611	2915	3165	2851	2940	2977	2893
	$\sigma$	148	114	237	223	133	161	182
	$\nu$	0,08	0,04	0,07	0,08	0,05	0,05	0,06

eliminam apenas os detalhes irrelevantes. Esta representação deve ser compacta de forma a promover maior eficiência computacional. Vários parâmetros vêm sendo utilizados, dos quais destacam-se: os coeficientes LPC, PARCOR, LSP e os MFCC. O objetivo deste estudo é identificar qual dos coeficientes citados apresenta melhor robustez quando são utilizados locutores diferentes na etapa de reconhecimento da fala. A seguir será feito um breve comentário a respeito de cada coeficiente mencionado e serão analisados os resultados obtidos.

#### A. Coeficientes LPC e Coeficientes de Reflexão PARCOR

Os coeficientes LPC e os coeficientes Parcor são obtidos através da análise por predição linear. O sinal acústico de voz pode ser modelado como a saída de um filtro linear variante no tempo excitado por um trem de pulsos para a fala sonora ou por um ruído branco para os sons surdos. A análise de predição linear consiste em se estimar do sinal de fala os coeficientes de predição denominados coeficientes LPC. O preditor linear, um filtro de ordem  $p$ , produz uma soma ponderada das  $p$  últimas amostras na entrada do preditor. A saída do filtro linear no  $n$ -ésimo instante da amostragem é dada por

$$s_n = Gu_n - \sum_{k=1}^p a_k s_{n-k}, \quad (1)$$

onde:

- $s_n$  – Sinal predito;
- $Gu_n$  – Termo de excitação;
- $a_k$  – Coeficientes preditores.

Desta forma temos que os coeficientes LPC correspondem aos coeficientes  $a_k$  da função de transferência do modelo de produção da voz representado por

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} = \frac{1}{A(z)}. \quad (2)$$

Os coeficientes são estimados diretamente do sinal da fala a partir de pequenos segmentos do sinal. A abordagem básica

consiste em encontrar um conjunto de coeficientes de predição que minimize o erro quadrático médio sobre um pequeno segmento do sinal. Os coeficientes LPC podem ser obtidos através do método das autocorrelações conforme detalhado em [4].

Observe que, para se calcular os coeficientes de predição de ordem  $p$ , é necessário determinar os coeficientes de predição de todos os preditores de ordem menor que  $p$ . As variáveis intermediárias são chamadas de coeficientes de reflexão, ou PARCOR (coeficientes de correlação parcial). Esses parâmetros caracterizam, assim como os coeficientes LPC, o preditor de forma única, apresentando a vantagem de oferecer melhores características de quantização e de interpolação em relação aos coeficientes LPC. O termo "coeficientes de reflexão" deve-se a uma interpretação física que surge quando se modela o trato vocal por uma sucessão infinita de tubos de diâmetros variáveis [4].

### B. Coeficientes LSP

Pode-se definir um novo conjunto de parâmetros no domínio da frequência equivalente ao conjunto de coeficientes de predição linear citados anteriormente. Este novo conjunto é conhecido como coeficientes LSP. A principal vantagem no uso destes coeficientes se deve também ao fato dos mesmos oferecerem melhores características de interpolação que os coeficientes LPC [5].

### C. Coeficientes Mel-Cepstrais

Modelos auditivos têm ajudado a melhorar o desempenho de reconhecedores de fala. O Mel é uma unidade de medida de percepção do pitch de um tom. Não corresponde linearmente a frequência física de um tom, pois o sistema auditivo humano não percebe linearmente o pitch ou a frequência de forma linear, em toda a faixa audível do espectro. Stevens e Volkman (1940) realizaram diversos experimentos e definiram a escala *Mel* através de um mapeamento entre a escala de frequência em (Hz) e a escala de frequência percebida (*Mel*). Este mapeamento é linear até aproximadamente 1 KHz e logarítmico para frequências superiores. A escala de frequência denominada *Mel* consiste numa aproximação da filtragem de banda crítica do ouvido humano. A partir da frequência-mel são obtidos os coeficientes Mel-Cepstrais (MFCC) conforme detalhado em [6].

### D. Análise em Componentes Principais (PCA)

A análise das Componentes Principais consiste numa transformação linear de " $m$ " variáveis originais em " $m$ " novas variáveis, de tal modo que a primeira nova variável computada seja responsável pela maior variação possível existente no conjunto de dados, a segunda pela maior variação possível restante, e assim por diante até que toda a variação do conjunto tenha sido explicada. O objetivo desta técnica neste caso consiste em permitir uma redução da dimensão dos dados, minimizando desta forma o erro.

Todas as gravações realizadas foram segmentadas em pequenos intervalos e extraídos os coeficientes LPC, PARCOR, LSP e MFCC. Para cada tipo de parâmetro utilizado foram obtidos 10 coeficientes de forma que o conjunto de dados passou a ser representado por um vetor de dimensão  $N \times 10$ , onde  $N$  corresponde ao número de frames analisados.

$$x_n = [x_1(n)x_2(n)...x_{10}(n)]^T, \quad (3)$$

onde  $[\cdot]^T$  denota a transposta da matriz. A variável  $x_n$ ,  $1 \leq n \leq N$  representa apenas um frame do sinal de áudio

analisado. O conjunto de todos os frames ( $M$ ) é representado pelos  $M$  vetores e pode ser representado por

$$X = [x(1)x(2)...x(M)], \quad (4)$$

onde cada coluna de  $X$  representa os 10 coeficientes de cada frame do sinal. A partir de então determina-se a matriz de covariância definida por

$$C = [X - \mu][X - \mu]^T, \quad (5)$$

onde  $\mu$  representa o vetor médio onde estão incluídas as médias ao longo de cada linha de  $X$ . Em seguida utiliza-se *decomposição em valores singulares (SVD)* [5] para se expressar a matriz de covariância da seguinte forma

$$C = USU^T, \quad (6)$$

Onde  $U$  é uma matriz cuja coluna são os autovetores de  $C$  e  $S$  é uma matriz diagonal contendo os respectivos autovalores de  $C$ .

A soma dos autovalores representa a variância total observada em  $C$ . Sendo assim, se a soma dos  $k$  primeiros autovalores atingirem uma proporção, como a que foi considerada, acima de 98% em relação a soma de todos os autovalores temos que os  $k$  primeiros autovetores da matriz  $C$  irão responder por aquela variância total observada no conjunto de dados. Para ilustrar o mencionado pode-se verificar na Figura 3 a recuperação da variância dos coeficientes LPC de um sinal acústico em função do número de componentes principais.

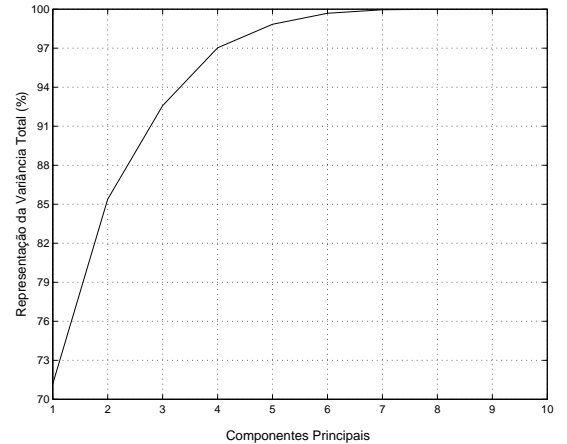


Figura 3. Recuperação da variância total dos coeficientes LPC da vogal /a/ pronunciada por 8 locutores.

Note que as 5 primeiras ( $k$ ) componentes representam mais de 98% da variância observada. Em suma, foi considerado um valor de  $k$ , que representa mais de 98% da variância total.

Utiliza-se então uma matriz formada pelas  $k$  primeiras colunas de  $U$  para definir a transformação linear

$$P = U_k^T (X - \mu), \quad (7)$$

Onde  $U_k$  representa uma matriz formada pelas primeiras  $k$  colunas de  $U$ . A matriz  $P$  corresponde aos coeficientes das componentes principais.

A equação acima nos permite representar os nossos dados por uma matriz  $P$  com dimensão menor que a original. Na seção seguinte serão apresentados os resultados e considerações obtidas após a extração dos diversos coeficientes mencionados e a redução do espaço dimensional dos mesmos através da utilização somente das componentes principais.

### E. Análise e Classificação de Padrões

Deve ser lembrado conforme citado anteriormente que uma ótima normalização é caracterizada por uma mínima variância intraclases e uma máxima variância interclasses. Para esta análise foi utilizado a medida de discriminabilidade ( $F$ ) que é determinada pela relação entre a variabilidade interclasse e a variabilidade intraclases, conforme representado abaixo. Nota-se que quanto maior esta relação, maior é a discriminância apresentada pelos dados.

$$F = \frac{\text{Variabilidade interclasse}}{\text{Variabilidade intraclasse}} = \frac{(V \frac{1}{N} \sum_{i=1}^N \bar{f}_i)}{\frac{1}{N} \sum_{i=1}^N V(f_i)}, \quad (8)$$

onde:

$F$  – Medida de Discriminabilidade;

$V$  – Função de Variância;

$N$  – Número de classes;

$\bar{f}_i$  – É a média dos dados característicos da enésima classe  $i$ .

#### 1) Análise das Medidas de Discriminabilidade dos Coeficientes Considerados

TABELA II

DISCRIMINABILIDADE PARA UM CONJUNTO DE 4 HOMENS E 4 MULHERES

Coeficientes	Discriminabilidade
LPC	1,21
LSP	1,63
Parcor	1,68
Mel-Cepstrais	2,45

TABELA III

DISCRIMINABILIDADE CONSIDERANDO APENAS UM LOCUTOR FEMININO

Coeficientes	Discriminabilidade
LPC	1,71
LSP	3,40
Parcor	2,62
Mel-Cepstrais	5,82

TABELA IV

DISCRIMINABILIDADE PARA GRUPO DE 4 LOCUTORES FEMININOS

Coeficientes	Discriminabilidade
LPC	1,25
LSP	1,79
Parcor	1,69
Mel-Cepstrais	2,93

TABELA V

DISCRIMINABILIDADE PARA GRUPO DE 4 LOCUTORES MASCULINOS

Coeficientes	Discriminabilidade
LPC	1,28
LSP	1,84
Parcor	1,46
Mel-Cepstrais	2,68

## IV. CONCLUSÕES

De acordo com os resultados obtidos observa-se que as frequências dos formantes se alteram ao longo do tempo e estas alterações são maiores para grupos de pessoas com características distintas (e.g. homens e mulheres). Algumas técnicas de normalização usam como referência o terceiro formante, pois como foi visto é o formante que apresenta menor variação ao longo do tempo.

Foi verificado que os coeficientes com melhor relação intra e extra classes são os Mel Cepstrais que apresentam maiores índices de discriminabilidade. Desta forma, pode-se considerar os coeficientes Mel-Cepstrais como sendo os mais robustos à variabilidade de locutor.

Outra consideração importante refere-se ao grande aumento dos índices de discriminabilidade obtidos quando analisa-se somente um locutor. Neste caso, foi constatado que os coeficientes Mel-Cepstrais apresentam uma relação de discriminabilidade bem superior aos demais coeficientes. Quando analisa-se o grupo feminino e o masculino observa-se, na maioria dos casos, um aumento também da discriminabilidade.

Na etapa de extração dos formantes muitas dificuldades foram encontradas por uma série de fatores já mencionados anteriormente. Onde observa-se que muitas vezes o próprio ouvido humano pode classificar vogais que não correspondem com as frequência de formantes medidas, induzindo ao erro de classificação.

A normalização de locutores é uma das principais áreas em reconhecimento de voz, pois com o emprego desta técnica a taxa de erros pode-se ser reduzida consideravelmente.

## REFERÊNCIAS BIBLIOGRÁFICAS

- [1] L. J. Gerstman, "Classification of Self-Normalized Vowels," *IEEE Transactions on Audio and Electroacoustics*, vol. Au-16, no. 1, pp.78-80, March 1968.
- [2] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *The Journal Of The Acoustical Society Of America*, vol. 24, no. 2, pp.175-184, March 1952.
- [3] S. F. Disner, "Evaluation of Vowel Normalization Procedures," *The Journal Of The Acoustical Society Of America*, vol. 67, no. 1, pp.253-261, January 1980.
- [4] J. R. Deller, J. H. L. Hansen and J. G. Proakis, *Discrete-Time Processing of Speech Signal*, IEE Press, 2000.
- [5] A. V. Barbosa, *Codificação Audiovisual Integrada da Fala*, Dissertação Mestrado-Universidade Federal de Minas Gerais, 2000.
- [6] H. P. Combrinck and E. C. Botha, "On The Mel-scaled Cepstrum," *Proceedings of the Seventh Annual South African Workshop on Pattern Recognition*, University of Cape Town, November 1996.
- [7] P. Zhan and M. Westphal, "Speaker Normalization based on Frequency Warping," *Proc. ICASSP-97*, Vol. 1, pp. 1039-1042, Munich, 1997.
- [8] E. Gouvêa and R. Stern, "Speaker Normalization through Formant-Based Warping of the Frequency Scale," *Eurospeech-97*, Vol. 3, pp. 1139-1142, Rhodes, 1997.
- [9] P. Boersma and D. Weenink, Praat is a freely available tool for speech analysis, developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences of the University of Amsterdam, The Netherlands. <http://www.fon.hum.uva.nl/praat>.