Compressão dos Coeficientes LSF via Modelo Neural

Dirceu Gonzaga
Instituto Militar de Engenharia
Depto. Engenharia Elétrica
Praça Gal. Tibúrcio, 80
Rio de Janeiro, RJ
22.290-270, Brasil
dirceu@epq.ime.eb.br

Antônio Carlos G. Thomé
Universidade Federal do Rio de Janeiro
DCC/NCE
P.O. Box 68504
Rio de Janeiro, RJ
21.945-970 Brasil
thome@nce.ufrj.br

Resumo— **Neste artigo é analisado um modelo neural para a compressão dos coeficientes LSF, os quais são utilizados na compressão de voz. É feita uma comparação entre a transformada de Karhunen Loève e a utilização do modelo neural auto-associativo com cálculo do erro de treinamento ponderado por um fator que insere características dos coeficientes. Este treinamento reduz a distorção espectral com maior eficácia. É apresentado, também, resultados da compressão com uma modelo neural auto-associativo não-linear, o qual apresentou o melhor resultados de todos os modelos estudados.**

I. Introdução

A compressão ou codificação de voz tem sido uma importante área de pesquisa há várias décadas, com um substancial progresso nas aplicações dos codificadores de voz a baixas taxas, que se encontra em franca evolução. Especialmente, nos últimos dez anos, a área vem experimentando um grande interesse motivado pela ampliação das aplicações em telecomunicações e armazenamento de voz.

Duas abordagens distintas podem ser usadas para a representação digital de sinais de voz: a representação por formas de onda e a representação paramétrica [6], [12]. A primeira procura simplesmente preservar a forma de onda do sinal de voz através de um processo de amostragem e quantização.

Na representação paramétrica, utiliza-se um modelo de produção da voz cujos parâmetros são ajustados para reproduzir com a maior fidelidade possível o intervalo de voz em análise. Este tipo de representação é utilizada para codificar a voz a taxas inferiores a 16 kbps. Existem dois parâmetros principais que são normalmente ajustados, os quais são: de excitação (relacionados à fonte geradora do sinal de excitação do aparelho vocal) e parâmetros do trato vocal (relacionados com a resposta em freqüência do aparelho vocal humano). Dentre os codificadores desse tipo encontram-se os codificadores CELP e os VOCODERS.

Uma das ferramentas mais poderosas para a análise da voz é o método da predição linear. Ela provê boas estimativas dos parâmetros da voz e é altamente eficiente do ponto de vista computacional (velocidade) [12]. Para a transmissão de voz, vários métodos já foram criados [12], mas com a introdução dos coeficientes LSF, por Itakura [9], que nada mais é do que uma forma de representar os coeficientes LPC, houve um avanço substancial em tais codificadores.

Na área de compressão sempre se buscou algoritmos com a finalidade de reduzir a redundância dos dados, para isso uma ferramenta considerada ótima é a transformada de Karhunen Loève (KLT) [10], [15] ou sua aproximação pela transformada cosseno [4]. Esses algoritmos realizam o que é conhecido como análise de componentes principais (PCA) e têm sido bastante utilizados em compressão de imagem [10], [14]. Recentemente tem crescido a utilização de redes neurais para realizar a mesma tarefa, tanto de forma não supervisionada [8] como através de redes auto-associativas [1]. Na compressão de voz pode-se utilizar os algoritmos diretamente sobre a forma de onda ou sobre parâmetros da fala, o que foi realizado em [5], [15] na compressão de espectro da voz e mais recentemente em [7], [13] na compressão dos coeficientes *LSF*.

Como a compressão de voz está ligada à obtenção de uma representação digital compacta do sinal de voz para uma transmissão ou armazenamento eficiente, o objetivo deste trabalho é descrever um estudo da capacidade de compressão dos coeficientes LSF por Redes Neurais Artificiais Auto-associativas, determinando a melhor topologia, normalização dos dados e forma de treinamento para uma compressão dos coeficientes, sem a preocupação com a codificação ou quantização dos mesmos.

Este artigo está organizado da seguinte forma: seção II é feita uma descrição dos coeficientes LSF e de suas principais característicasna, na seção III são apresentadas as definições da análise de componentes principais, em seguida na seção IV é feita uma descrição das redes neurais auto-associativas aplicadas à compressão de dados. Na seção V são apresentados os resultados obtidos na compressão dos coeficientes LSF e seus efeitos no espectro de voz. Finalmente na seção VI é feita uma conclusão onde destacam-se os principais resultados da pesquisa.

II. Coeficientes LSF

Os coeficientes LSF (*Line Spectrum Frequency*) constituem uma das várias representações possíveis para os coeficientes de predição α_i do filtro de síntese utilizado na

TABELA I

VALORES TÍPICOS DOS COEFICIENTES LSF

Coeficiente	Máximo	Mínimo	Média	Variância
1	0.4413	0.0460	0.1787	0.0032
2	0.7773	0.1183	0.3222	0.0156
3	1.1313	0.2346	0.5316	0.0338
4	1.4556	0.4774	0.8753	0.0359
5	1.7652	0.6131	1.2395	0.0459
6	2.0829	0.9524	1.5505	0.0380
7	2.3123	1.4554	1.9238	0.0164
8	2.5941	1.7957	2.1938	0.0134
9	2.8351	2.1547	2.5268	0.0088
10	3.0533	2.4658	2.8155	0.0059

análise LPC. Este filtro é definido por:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} \alpha_i z^{-i}} \tag{1}$$

onde $A(z)$ é o filtro inverso de ordem p.

Para o cálculo dos coeficientes *LSF* é necessário definir dois polinômios auxiliares $P(z)$ e $Q(z)$ obtidos a partir de $A(z)$ da seguinte forma [9]:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{2}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{3}$$

onde, $P(z)$ é um polinômio simétrico e $Q(z)$ é um polinômio antissimétrico.

As raízes de $P(z)$ e $Q(z)$ determinam os coeficientes *LSF*. Estes polinômios possuem ligação direta com o modelo acústico do trato vocal e com os estágios do filtro preditor com estrutura em treliça.

Se H(z) é estável, então:

- as raízes de $P(z)$ e $Q(z)$ estão sobre o circulo unitário.
- As raízes de $P(z)$ estão alternadas com as raízes de $Q(z)$, ou seja, $r_1 < q_1 < r_2 < q_2 < \cdots < q_{p+1}$, onde r_i e q_i representam a posição angular da i-ésima raiz de $P(z)$ e $Q(z)$, respectivamente.
- O filtro $H(z)$ continuará estável após a quantização das raízes de $P(z)$ e $Q(z)$ desde que os dois itens anteriores sejam respeitados pelos valores quantizados.

Além disso os coeficientes LSF possuem uma faixa dinâmica bem comportada, possibilitando uma quantização mais eficiente, do que as outras formas de representação dos coeficientes LPC.

Finalmente, os coeficientes de predição de $A(z)$ são obtidos a partir dos coeficientes de $P(z)$ e $Q(z)$ através da seguinte igualdade polinomial:

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{4}$$

A. Características Estatísticas dos LSF

A fim de conhecer um pouco sobre a capacidade de compressão dos coeficientes LSF, fez-se um estudo prévio das suas características estatísticas.

No item II, foi visto que uma das característica dos LSF é a sua ordenação, o que pode ser constatado na Tabela I, que mostra os valores máximos, mínimos, as médias e as variância de cada coeficiente. Pode-se ver que a distância entre os primeiros e os últimos coeficientes é maior que uma ordem de grandeza. Se for feita uma comparação entre dois vetores de coeficientes, utilizando-se a distância euclidiana, a diferença será dominada pelos termos de maior valor absoluto, enquanto que a informação mais importante encontra-se nos coeficientes de menor valor. Este é uma fato que deve ser levado em conta na utilização de uma rede neural com o aprendizado pelo erro médio quadrático, pois a rede tenderá a aprender com mais precisão os coeficientes de maior variabilidade.

III. ANÁLISE DOS COMPONENTES PRINCIPAIS

No processo de análise dos componentes principais, o objetivo é reduzir a dimensionalidade dos dados através da descorrelação dos coeficientes, girando os eixos coordenados até que a matriz de covariância se transforme em uma matriz diagonal. Esta transformação é linear e representada como:

$$C_\nu = \Phi \Lambda \Phi^T \tag{5}$$

onde C_ν é a matriz de covariâncias de ν, Φ é a matriz de autovetores e Λ a matriz de autovalores. O sobrescrito T é de transposto. Cada elemento de C_ν, ou seja, $C_\nu(i,j)$ é calculado por:

$$C_\nu(i,j) = \frac{1}{N_q} \sum_{m=0}^{N_q - 1} (\nu_m(i) - \mu_\nu(i))(\nu_m(j) - \mu_\nu(j)) \tag{6}$$

onde N_q é o número de quadros de voz i e j são os índices dos coeficientes e μ_ν é o vetor média dos coeficientes.

Resolvendo a Eq. 5, obtemos os autovetores e autovalores da transformação. Os autovalores fornecem dados importantes sobre a quantidade de informação do sinal. Sendo N_v o número de autovetores do sistema, e tomando todos os autovalores $\lambda_0, \lambda_1, \cdots, \lambda_{N_v-1}$ em ordem decrescente, obtém-se uma importante relação dada por [16]:

$$\sum_{i=0}^{N_v-1} \lambda_i = Tr(C_\nu) = \sum_{i=0}^{N_v-1} \sigma_i^2 \tag{7}$$

onde Tr é o traço de matriz e σ_i é a variância do i-ésimo coeficiente. Esta equação mostra, também, que os autovalores estão diretamente relacionados com a variância do sinal

Busca-se com esse procedimento realizar uma transformação para um espaço, cuja base é formada pelo conjunto de autovetores. Nestas novas dimensões os autovalores são as variâncias dos dados transformados.

Pode-se definir a quantidade de informação contida em cada autovalor com o valor obtido pela expressão:

$$\zeta_i = \frac{\lambda_i}{\sum_{i=0}^{N_v-1} \lambda_i} * 100 \quad \% \tag{8}$$

e a percentagem total de informação acumulada pelos primeiros autovalores (uma dimensão corresponde a um autovalor/ autovetor) por:

$$R_m = \frac{\sum_{j=0}^{m-1} \lambda_i}{\sum_{i=0}^{N_v-1} \lambda_i} * 100 \quad \% \tag{9}$$

onde m é o número de dimensões desejadas.

A Figura III apresenta um gráfico dos valores das dimensões acumuladas dada pela Eq. 9, onde se vê que o primeiro autovalor corresponde a cerca de 53% da informação original do sinal e considerando os 5 maiores autovalores, esse percentual sobe para 90%.

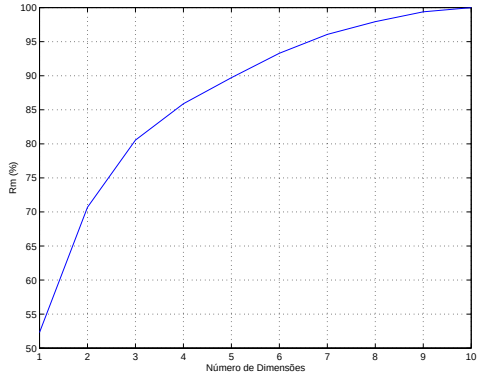

Fig. 1
Gráfico de variâncias acumuladas

IV. Redes Neurais Auto-Associativas

A compressão de dados é bastante utilizada em processamento de imagem, através do uso de transformadas cosseno, que podem servir como uma aproximação das transformadas de Karhunen Loève (KLT) [4]. O esquema da aplicação de transformadas na codificação é descrito na Figura 2, onde é apresentado um conjunto de dados ou coeficientes a serem transmitidos, aplicando-se em seguida uma transformada T, depois procede-se com a quantização dos dados transformados. No receptor aplica-se o processo inverso. Com esse processo procura-se extrair os

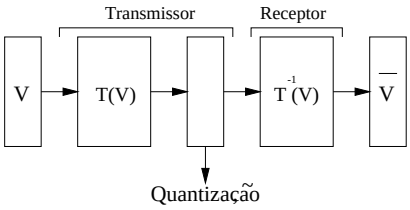

Fig. 2
Esquema de compressão utilizando transformadas

componentes principais do sinal, pois estes concentram a informação útil do sinal. A análise de componentes principais pode ser feita através do algoritmo *Karhunen Loève Transform* (KLT) ou de redes neurais fazendo *Principal Component Analisys* (PCA) [8], [11]. Contudo o algoritmo KLT é inviável para ser implementado em tempo real visto que necessita de armazenamento de dados e das estatísticas do sinal, por isso utilizam-se algoritmos sub-ótimos como as transformadas cosseno. A redes PCA, que são redes não supervisionadas, aproximam as KLT, onde os pesos tendem aos autovetores do sinal [8].

Uma outra forma de achar os componentes principais é utilizando redes do tipo *Multi Layer Perceptron* Auto-Associativas (MLPAA) com função de transferência linear, onde a camada intermediária tem um papel importante na atuação como um detetor das características principais [8], [11]. Uma das características deste modelo de extração de características sobre o modelo PCA convencional (treinamento hebbiano não supervisionado) está na possibilidade de se introduzir algum conhecimento sobre o problema ou fenômeno no processo de treinamento da rede neural.

A Figura 3 apresenta uma possível forma de emprego das redes neurais MLP para compressão e transmissão de dados. A rede neural apresentada utiliza uma estrutura simétrica com funções lineares, que após o treinamento pode ser seccionada em duas partes: uma localizada no transmissor, a fim de proceder a compressão e quantização e a outra no receptor, fazendo a descompressão. Um rede MLP operan-

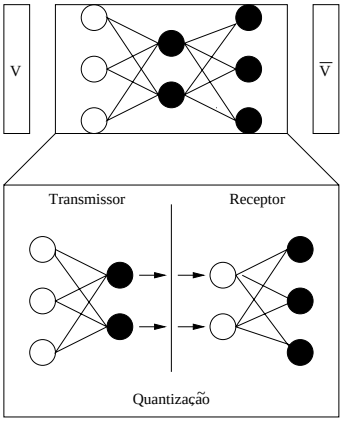

Fig. 3
Esquema de compressão e transmissão utilizando Redes Neurais. Os círculos cheios são as camadas com função de transferência linear ou não linear

do desta forma é referenciada como *autoencoder* ou *autoassociator*[8]. Segundo a literatura, e que se comprovado na prática, tais estruturas calculam os componentes principais dos dados de entrada, que oferecem uma base ótima para redução linear da dimensão dos dados [8]. O conhecimento físico do problema pode ser inserido durante o treinamento, como por exemplo, na compressão do coeficientes LSF é conhecido [3] que os primeiros coeficientes têm mais importância que os últimos e assim pode-se utilizar este conhecimento na aplicação do algoritmo de aprendizado.

Segundo [8] a introdução de neurônios não lineares (sigmóide ou tangente hiperbólica) possibilitam a extração de estatísticas de mais alta ordem, acrescentando mais robustez para expansão.

Para demostrar o efeito do PCA, a Figura 4 mostra a estimação dos dados feita por um PCA com compressão de dados de duas para uma dimensão, ou seja, uma MLPAA com a configuração de um neurônio na camada interme-

diária e dois na saída. Pode-se perceber que os dados estimados acompanham a dimensão de maior variância dos dados originais. Neste trabalho será adotada a seguinte padronização: A camada de entrada será chamada de nE, onde n é a dimensão do vetor de entrada; as demais camadas serão chamadas de kL/N, onde k é a dimensão da camada, ou o número de neurônios da camada e L/N faz referência a função de transferência (L)inear ou (N)ão-Linear. Para dados com uma distribuição mais complexa, o PCA

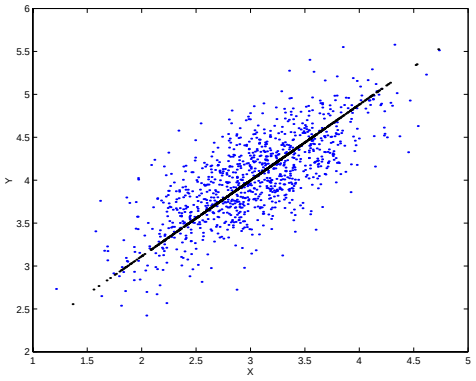

Fig. 4

Exemplo dos efeitos do PCA linear, para dados bem comportados. A linha por cima dos pontos é a estimação feita pelo PCA após a descompressão utilizando-se uma rede com 2E-1L-2L neurônios

não pode fazer uma boa estimação da dimensões principais, conforme pode ser visto na Figura 5.

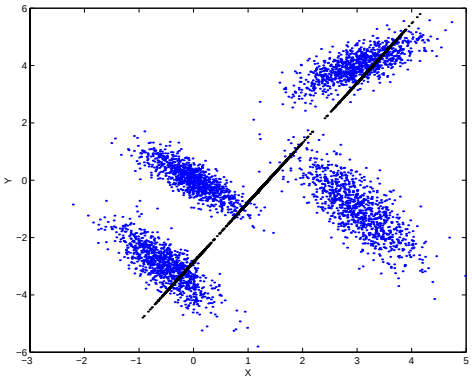

Fig. 5

Estimação feita pelo PCA, mostrando que uma rede linear não consegue aproximar adequadamente os dados. Rede AA com 2E-1L-2L neurônios

Com isso a adição de camadas escondidas, conforme a Figura 6, permite ao *autoencoder* desenvolver uma compressão não linear dos dados, potencialmente com resultados mais efetivos.

A Figura 7 mostra uma distribuição de dados com 4 agrupamentos distintos com direções principais distintas,

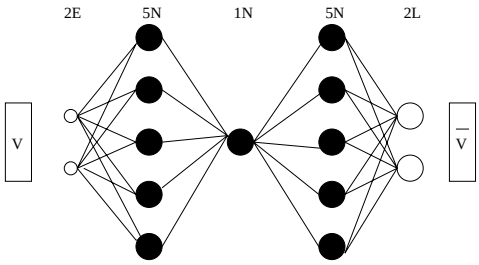

Fig. 6

Esquema de compressão não linear com camadas escondidas

que serão utilizados para mostrar as deficiências do PCA e a eficiência da rede não linear.

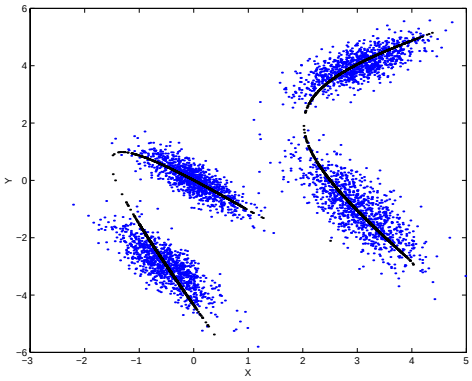

Fig. 7

Estimação feita pelo PCA não-linear, mostrando que esta rede já consegue acompanhar as direções de maior variância do sinal. Foi utilizada uma MLPAA com 2E-5N-1N-5N-2L neurônios

V. Análise dos Resultados

Para os testes foi utilizada a base de dados descrita em [2], de onde foram extraídas 70 frases foneticamente balanceadas para treinamento e 30 para teste, sendo estas diferentes das primeiras, gravadas por 2 homens e 1 mulher. O sinal foi dividido em quadros de 25 ms, dando um total de 5000 vetores de coeficientes LSF para treinamento e 2000 para teste.

Utilizou-se a distorção espectral (DE) para medir a diferença espectral gerada pelos coeficientes LPC após a descompressão do LSF. A distorção espectral, em dB, é definida por:

$$DE = \left[\frac{1}{f_h - f_l} \sum_{j=f_l}^{f_h - 1} \left| 10log_{10}|S_j|^2 - 10log_{10}|\overline{S}_j|^2 \right| \right]^{1/2} \quad (10)$$

onde S_j e $\overline{S}_j$ são respectivamente, a potência espectral original e estimada, produzida pelos parâmetros LPC.

A Tabela II apresenta os resultados obtidos com a utilização do PCA através de uma MLPAA 10E-nL-10L, onde é n é o número de neurônios de compressão.

TABELA II

DISTORÇÃO ESPECTRAL MÉDIA EM DB COM O PCA

Nr de Dimensões	Dados	
	Treinamento	Teste
5	2.18	2.26
6	1.81	1.83
7	1.38	1.44
8	1.03	1.05
9	0.57	0.61

A Figura 8 mostra a distorção espectral média obtida com o PCA linear. Pode-se ver que as freqüências mais baixas e as mais altas tiveram os maiores valores de distorção, visto que os dois primeiros coeficientes e os dois últimos apresentam as menores variâncias, fazendo com que a rede dê maior importância aos coeficientes intermediários.

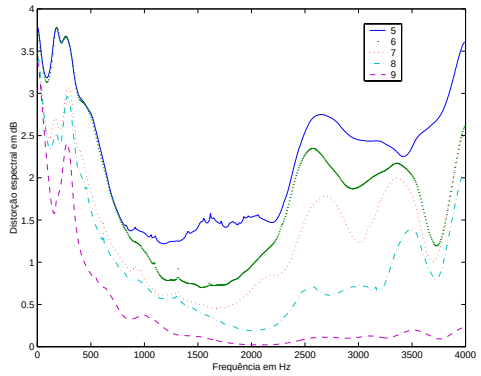

Fig. 8

DISTORÇÃO ESPECTRAL MÉDIA PARA O PCA LINEAR

Em [3] encontra-se descrita uma função de distância, que tem sido utilizada em muitos trabalhos de quantização vetorial para voz, pois ela leva em conta características do espectro de freqüências de voz que são desprezadas na distância euclidiana. A função é descrita da seguinte forma:

$$d(\vec{a},\vec{\overline{a}}) = \sum_{i=1}^{p}[w_i(a_i - \overline{a}_i)]^2 \qquad (11)$$

onde w pode ser descrito de várias formas. O w escolhido neste trabalho é dado pela equação:

$$w_i = \frac{1}{a_i - a_{i-1}} + \frac{1}{a_{i+1} - a_i} \qquad (12)$$

onde a_i é o i-ésimo coeficiente LSF. Esse peso ressalta os coeficientes LSF de maior importância espectral, que são aqueles correspondentes aos formantes do trato vocal.

Realizando o treinamento da RNA através da Eq 11, pode-se introduzir o conhecimento do problema ao treinamento da rede, reduzindo dessa forma a distorção espectral conforme pode ser visto na Tabela III.

Todos os resultados tiveram seus valores reduzidos visto que a rede conseguiu aprender os coeficientes LSF que

TABELA III

DISTORÇÃO ESPECTRAL MÉDIA EM DB COM O PCA PONDERADO

Nr de Dimensões	Dados	
	Treinamento	Teste
5	2.14	2.25
6	1.77	1.83
7	1.37	1.43
8	1.01	1.04
9	0.58	0.62

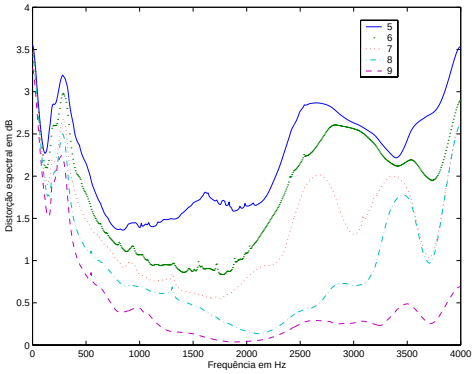

Fig. 9

DISTORÇÃO ESPECTRAL MÉDIA PARA O PCA LINEAR PONDERADO

contribuíam com a maior potência espectral no cálculo da distorção. Este efeito pode ser visto na Figura 9. A fim de verificar as não-linearidades dos coeficientes LSF, foram realizados treinamentos com um PCA não-liner, conforme mostrado na Tabela IV. Pode-se perceber que os resultados foram ainda melhores que os do PCA linear, o que mostra que os LSF possuem não linearidades que foram mapeadas pela rede não linear. Pode-se pensar nesse resultado com o do exemplo apresentado na seção IV. A Figura 10 apresenta a distorção epsectral obtida com o PCA não-linear, mostrando que os valores de distorção devido às ferqüências intermediários baixaram enquanto as freqüências baixas e altas permaneceram no níveis do PCA linear.

TABELA IV

DISTORÇÃO ESPECTRAL MÉDIA EM DB COM A MLPAA NÃO LINEAR

Nr de Dimensões	Dados	
	Treinamento	Teste
20N-5N-20N-10L	2.01	2.13
17N-6N-17N-10L	1.66	1.73

VI. CONCLUSÃO

Neste artigo analizou-se um modelo neural para compressão dos coeficientes LSF, coeficientes esses utilizados na compressão de voz. A utilização de redes neurais auto-

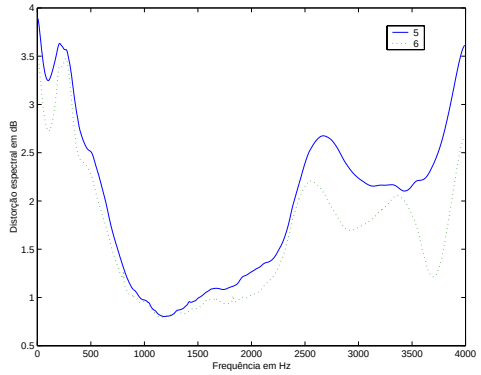

Fig. 10
DISTORÇÃO ESPECTRAL MÉDIA PARA O PCA NÃO-LINEAR

associativas mostram ser mais eficientes que a KLT, visto que com a rede é possível se controlar o treinamento, introduzindo um conhecimento específico sobre os dados. As rede auto-associativas não-lineares apresentam, no caso, um melhor desempenho que as rede lineares. O próximo passo nessa pesquisa é verificar o efeito da quantização dos novos coeficientes na saída do transmissor e realizar algumas medidas subjetivas da qualidade da voz sintetizada no receptor.

REFERÊNCIAS

[1] Abidi, M.A., Yasuki, S., Crilly, P.B. *Image Compression using Hybrid Neural Networks combining the auto-associative multilayer perceptron and the self-organizing feature map.* IEEE Transaction on Consumer Electronics, Vol. 40, n. 4, nov. 1994.

[2] Alcaim, A., Solewicz J.A., Moraes, J.A.. *Frequência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro* . Revista da Sociedade Brasileira de Telecomunicações, vol. 7, n. 1, dez. 1992.

[3] Atal, B.S.; Paliwal, K.K. *Efficient Vector Qunatization of LPC Parameters at 24 bits/frame.* ICASSP, p. 661-664, 1991.

[4] Clarke, R.J., Tech B. et al *Relation between the Karunen Loève and cosine transforms.* IEE Proc, vol. 128, Pt F, n. 6, nov. 1981.

[5] Crowther, W.R., Rader, C.M. *Efficient Coding of Vocoder Channel Signals Using Linear Transformation.* Proceedings of the IEEE, p.1594-1595, nov. 1966.

[6] Deller, J.R. JR, Proakis, J.G., Hansen, J.H.L.. *Discret Time Processing of Speech Signals.* Macmillan Publishing Company, New York, 1993.

[7] Farvardin N., Larola Rajiv *Efficient Encoding of Speech LSP Parameters Using the Discrete Cosine Transformation.* IEEE Proceedings ICASSP, p. 168-171, 1989.

[8] Haykin, S. *Neural Networks: A Comprehensive Foundation.* New York, Macmillan, 1994.

[9] Itakura, F. *Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals.* J. Acoustic Soc. America, vol 57, S35(A), 1975.

[10] Lim, Jae S., Lim, Joe S. *Two-Dimensional Signal and Image Processing.* Prentice Hall Inc., 1989.

[11] Oja, E. *Principal Components, Minor Components, and Linear Neural Networks* . Neural Networks, vol. 5, p. 927-935, 1992

[12] Rabiner, L. R., Shafer, R. W. *Digital Processing of Speech Signals.* Prentice-Hall Inc., 1978.

[13] Vu, Hai Le; Lois, Laszlo *Optimal Transformation of LSP Parameters Using Neural Network.* IEEE Proceedings ICASSP, p. 1339-1342, 1997.

[14] Wintz P. A. *Transform Picture Coding.* Proceedings of the IEEE, vol. 60, n. 7, p.809-819, 1972.

[15] Zahorian Stephen A. *Principal-components analysis for low-redundancy encoding of speech spectra.* J. Acoust. Soc. Am., vol. 63(3), p. 832-845, mar. 1981.

[16] Picone J.W. *Signal Modeling Techniques in Speech Recognition.* Proceedings of the IEEE, vol. 81, n.9, set. 1993.