Proceedings of V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 25-30, April 2-5, 2001- Rio de Janeiro – RJ -Brazil

Reconhecimento de Locutores em Língua Portuguêsa com Modelos de Redes Neurais e Gaussianos

André Machado Caricatti [1], Li Weigang[2]
[1] Instituto Nacional de Criminalística, Departamento de Polícia Federal, Brasil
[2] Departamento de Ciência da Computação – CIC
Universidade de Brasília – UnB, Brasília – DF, Brasil
E-mails: caricatti.amc@dpf.gov.br, weigang@cic.unb.br

Abstract

To study the speaker recognition problem, the mel-cepstral coefficients and their derivatives, delta-mel-cepstral coefficients are used as the classification parameters. A database is formed from 14 speakers in a closed environment using Portuguese with noise an echo. To execute text-independent speaker recognition, self-organizing maps (SOM) and Gaussian Mixture Models (GMM) are used together with relative entropy calculations among test and reference speaker models. Specially in the GMM application, 20 seconds is enough to form the models with 32 mixtures per speaker, and the correct verifications arrives at a rate of 64 % in the best case.

1. Introdução

Para o reconhecimento automático de locutores vislumbram-se diversas aplicações, merecendo destaque o emprego na área de segurança, ao permitir controles biométricos de autenticação e acesso, seja a sistemas computacionais, espaços físicos e outros. Na função de perito criminal federal, a motivação para o aprofundamento dos estudos foi a possibilidade de realizar a verificação de locutores para determinar autoria de gravações, fundamentais na elucidação de crimes. São poucos os estudos desenvolvidos sobre locutores de língua portuguesa, menos ainda para brasileiros. A literatura revela que a maioria dos estudos procura a melhoria de performance de sistemas que operem, principalmente, com amostragem de sinais a partir de aparelhos telefônicos. Porém, as situações reais nas quais se obtém amostras para formação de provas em procedimentos judiciais são as mais adversas, e faltam experimentos nestas condições. Adicionalmente, surge a possibilidade de emprego da tecnologia para, por exemplo, separar as gravações, por autor, durante realizações de interceptações de fluxo de comunicação telefônica. Embora pareça uma tarefa simples, o grande volume de casos e os longos períodos de apuração a transformam em verdadeiros empecilhos no desenrolar de investigações.

2. Coleta de Dados

Um grupo de 14 pessoas foi selecionado para produzir locuções previamente determinadas.

Seguindo a orientação do estudo de fonética, é de se esperar que o ser humano, ao exprimir-se pela voz em diferentes datas, varie as características de longo tempo, a prosódia, tanto quanto as características de curto tempo, estas últimas empregadas para todas as implementações deste trabalho. Porém, para alcançar o que se procede naturalmente na obtenção de locuções que sirvam como vestígios para fins processuais criminais, normalmente padrões distribuídos imprevisivelmente no tempo, tratou-se de programar coletas em quatro datas.

Foram selecionadas 11 frases, listadas na tabela 1, contendo boa quantidade de fonemas nasalados e sons vocalizados [1]. As locuções foram então geradas pelos 14 locutores, em 4 seções dispostas ao longo de 4 meses. Em cada sessão, os locutores repetiram as frase 4 vezes.

Tabela 1 : Frases empregadas nos experimentos.

Número	Texto
1	O prazo ta terminado
2	Amanhã ligo de novo
3	Aguarde próximo contato
4	Ta tudo sob controle
5	Ta tudo bem
6	Fui caminhando até
7	O ponto de encontro
8	Não encontrei o material
9	Ele estava aguardando a entrega
10	A confiança surge quando
11	As atitudes são claras e seguras

Para contornar a expressão mecânica das pessoas ao repetir textos com encadeamento já conhecido, alternou-se a sequência de repetição. Sobre o ambiente de coleta, foi reservada a mesma sala, um auditório, para a obtenção das amostras, não havendo no recinto

nenhuma preparação especial destinada à redução de ecos ou eliminação total de ruído de fundo.

3. Metodologias de Reconhecimento

3.1 Parametrização do sinal

Extração de Parâmetros: para capturar a dinâmica dos movimentos do trato vocal, calcula-se o espectro do sinal, em geral por intervalos de 10 a 20 ms, seja por meio de coeficientes resultantes de análise de Fourier ou por valores extraídos de bancos de filtros. É comum também representar o envelope espectral indiretamente por parâmetros de um modelo de predição linear *"all-pole"*, ou em termos dos primeiros coeficientes *cepstrais* [2], que representam a transformação inversa da função logaritmo aplicada sobre o sinal transformado.

Análise por Predição Linear: os modelos descritivos de produção de locuções tentam descrever os diversos eventos que as compõem, partindo de sinais periódicos, como as vogais, indo até os ruídos quase aleatórios de fricativas [3].

O Cepstrum: ao isolar a locução composta por sinais periódicos, ou vocalizados, tem-se a geração de pulsos na glote que, passando pelo trato vocal, agindo como um elemento filtrante, resultam em vocóides. Uma equação que os definiria como convolução entre excitação *U* e filtro *H*, trazida para o domínio de frequências através da transformada de Fourier, seria:

$$S(w) = G\,U(w)\;H(w) \qquad (1)$$

Lembrando que o trato vocal possui características bastante individualizadas, enquanto nos pulsos que saem das cordas vocais é fácil encontrar semelhanças [4], toma-se por objetivo no reconhecimento de locutores a busca por uma separação entre estes componentes. Daí que a eq. (1) não vai ser de grande utilidade, pois fazê-lo é tarefa bastante complexa.

Como simplificação, deixando de lado a fase dos sinais, para que se tenha um sistema linear aditivo, aplicam-se logaritmos em ambos os lados, chegando a:

$$\log (S(w)) = \log (GU(w)H(w))$$

$$= \log (GU(w)) + \log (H(w)) \qquad (2)$$

Aplicando a transformada inversa de Fourier sobre a eq. (2) chega-se ao ***cesptrum*** real, ou cepstrum, uma troca de letras na palavra spectrum. Diz-se real porque a operação com logaritmo provoca a perda de informação da fase. O cálculo destes parâmetros pode ser feito a partir de modelagem dos sons vocalizados por predição linear, assumindo o sinal como sendo de fase mínima. Se assim for visto, encaixa-se o cepstrum no processamento homomórfico de sinais [5], que trata de transformar sinais combinados não-linearmente, para domínios lineares que permitam aplicação de técnicas convencionais, seguido de transformação inversa para o domínio original.

*As Escalas de Percepção Auditiva***:** uma razão para utilizar o espectro do sinal vem da observação do processo realizado na colclea do ouvido humano. Neste, o sinal sonoro é avaliado como se houvesse um banco de filtros, dispostos de forma não-linear. Porém, o ouvido humano é mais sensível para observar variações em frequências baixas, exigindo distâncias exponencialmente maiores para perceber alterações a medida em que caminha-se para as frequências altas, acima de 1 Khz [3,6].

3.2 Escolha de parâmetros

Um conjunto de parâmetros será suficiente, a medida em que cada um destes apresentar informação nova aos algoritmos de classificação. A partir do sinal de voz, podem ser retirados inúmeros destes, tornando a busca por conjuntos eficientes, se realizada de forma aleatória, inviável pelo crescimento exponencial de possibilidades.

Distribuição Normal: a distribuição normal, via de regra, é considerada uma boa aproximação para fenômenos do mundo real, capaz de capturar a variação de valores como parâmetros mel-cepstrais [3,7,8]. Além desta consideração, ela possui vantagens como ser facilmente parametrizável utilizando apenas as médias e variâncias. Adicionalmente, a soma de variáveis randômicas com distribuição normal produz uma nova variável randômica com distribuição normal.

Figura 1. Algoritmo Estimar-Maximizar *[7]*

Algoritmo Estimar - Maximizar

i. Distribuir os vetores de dados de entrada (parâmetros) entre *k* classes. O número *k* deve ser determinado préviamente. Esta distribuição pode ser feita aleatoriamente, ou de acordo com algum processo de agrupamento (*clustering);*
ii. Obter estimativas iniciais para as densidades, calculando médias e variâncias. A quantidade de vetores em cada classe pode servir como peso;
iii. Para cada vetor, estimar valores para cada uma das densidades do modelo;
iv. Selecionar a densidade que proporciona o maior valor, inserindo nela o vetor;
v. Se não houver qualquer mudança de classe, parar;
vi. Estimar novamente todas as densidades;
vii. Retornar ao passo iii.

Modelos de Misturas Gaussianas: um modelo de misturas gaussianas é composto por uma soma ponderada de densidades formulada pela eq. (3) [7].

$$p(x) = S\;P(w_i)\;p\,(x|w_i) \qquad (3)$$

$S\ P(w_i) = 1$
$P(w_i)$ *: peso, ou probabilidade "a priori" de* w_i
$p(x|w_i)$ *: probabilidade de x para a densidade* w_i

Para gerar uma observação a partir de um modelo de misturas, as funções de probabilidades de cada densidade do modelo são ponderadas e somadas. O treinamento destes modelos é realizado por algoritmos do tipo *estimar-maximizar (estimate-maximize)* [7].

Modelos de misturas gaussianas se assemelham a identificação por quantização de vetores, pois as locuções são modeladas por componentes agrupados (*clusters)*.

Uso de Redes Neurais: Neste trabalho, um modelo conexionista, ou *rede neural*, é comparado a outro modelo puramente estocástico. Vale ressaltar que outras modelagens, destacando-se as probabilísticas utilizando cadeias escondidas de Markov (HMM's), são igualmente relevantes como ferramentas de classificação de sinais de locução [8].

4. Implementação dos Modelos

O trabalho está inserido dentro do contexto de reconhecimento de padrões e, portanto, cabe a divisão dos sistemas em etapas de pré-processamento, extração de parâmetros e classificação.

No primeiro sistema, ao realizar classificações por meio de *mapas auto-organizáveis (SOM - self-organizing maps)* de Kohonen, buscou-se uma forma de, por meio da frequência de ocorrência de classes para cada locutor, obter informação que possibilitasse um trabalho independente do texto falado (Figura 2).

Já na segunda implementação, foram calculados, para cada locutor, misturas de modelos de misturas gaussianas, com a mesma intenção de utilizar uma técnica que permitisse independência de texto.

4.1 Descrição das etapas de processamento de sinais

Pré-processamento: com 3 etapas: amostragem, eliminação de pausas e pré-enfase.

- Fez-se uso de microfone de membrana, uma vez que sua resposta é equivalente para as principais frequências que compõe a voz, situadas na banda 10 Hz - 5 KHz. A taxa de amostragem de 11 KHz foi empregada, ou 5.500 Hz de Nyquist. Para a codificação das amplitudes utilizou-se PCM com 16 bits.
- De início, destaca-se trecho de locução a servir como modelo para pausa. Por meio deste, primeiramente são calculados a média e desvio padrão para cruzamentos de zero, bem como a média e desvio padrão para amplitude de onda. Com base nestes dados, são considerados trechos de "silêncio" aqueles cujas médias de cruzamentos de zero e amplitude de onda situam-se abaixo das médias, subtraídos destas os respectivos desvios padrões, obtidas para a pausa.
- A energia das frequências mais baixas tem destaque em relação as mais altas, resultando em parâmetros que, embora naturais, provocam distorções prejudiciais. Uma forma de contornar este obstáculo é empregar filtros que compensam este efeito, denominados filtros de pré-ênfase.

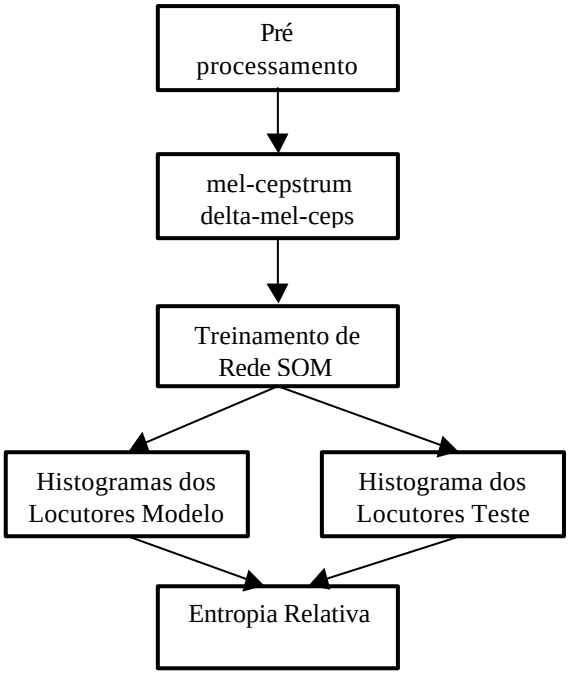
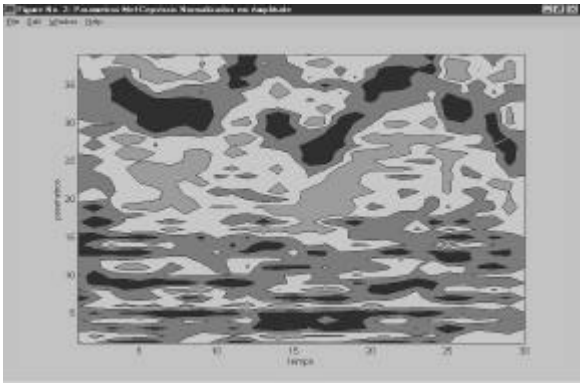


Figura 2. Esquema para um teste utilizando redes SOM.

Extração de Parâmetros da Locução: A escolha de coeficientes *mel-cepstrais* apenas acompanha a tendência predominante no reconhecimento de locutores. Os passos executados para obtê-los resumem-se em:

- calcular magnitudes reais no domínio das frequências, por meio de transformações de Fourrier;
- filtrar as magnitudes segundo a escala *mel*, em 40 bandas;
- aplicar logaritmos e transformada discreta de cosseno, gerando 40 parâmetros por janela .

Normalização das Amplitudes: Os parâmetros resultantes da etapa anterior são diretamente proporcionais aos níveis de energia dos sinais coletados, fator que pode deturpar a classificação indevidamente. Para fazer a normalização destas amplitudes, faz-se a divisão, a cada locução analisada, de cada um dos parâmetros por valores-limite, respectivos às diversas bandas. Assim, os parâmetros passam a se situar sempre na faixa de 0 a 1.



a

b

Figura 3. Gráficos de contornos para parâmetros mel-cepstrais retirados da locução “O prazo ta encerrado”:
a - sem normalização;
b - com normalização.

*Parâmetros Delta-Mel-Cepstrais***:** são calculadas as inclinações de parâmetros mel-cepstrais, segundo aproximação por polinômios, como proposta em [9]

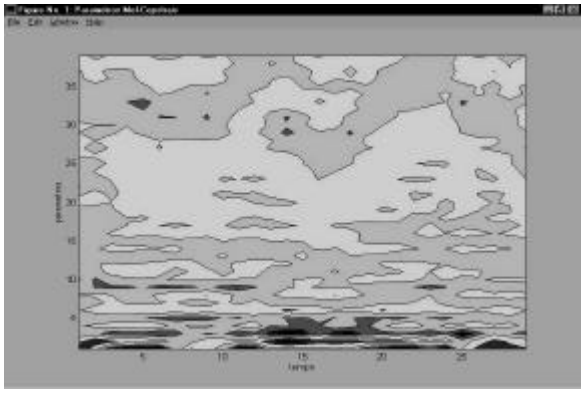
$$Dcep_L = \sum_{-k}^{k} -k\, cep_{L-k} \qquad k = 2; \qquad (4)$$

cep_L : vetor de parâmetros no instante L;

5. Experimentos

Durante as sequências de experimentos executados com redes SOM (9 experimentos) e modelos de misturas gaussianas (3 experimentos), a quantidade, em vetores de treinamento, a ser utilizada para o treinamento bem como para os testes, foi uma das principais incógnitas. É preciso lembrar que cada vetor de treinamento foi composto por dois grupos de valores, sendo eles:

- parâmetros mel-ceptrais, extraídos de janelas com aproximadamente 46 ms de locução, a partir de 512 pontos de amostragem a 11 khz , sucedendo-se a cada 12 ms, ou 128 pontos de amostragem;
- parâmetros delta-mel-ceptrais, calculados por aproximação polinomial de 5 janelas de parâmetros mel-cepstrais.



5.1 SOM e histogramas de ocorrência

Elementos que permaneceram inalterados nesta sequência de testes foram, a princípios, os parâmetros de entrada, compostos por 20 coeficientes mel-cepstrais e 20 coeficientes delta-mel-cepstrais, calculados por aproximação polinomial, de acordo com a eq. (4), computados em intervalos de 5 vetores, num total de 40 parâmetros.

O treinamento das redes SOM foi sempre realizado com taxa de aprendizado inicial igual a 1, limitando as iterações em 50 vezes o total de frases empregadas.

Relembrando o significado da entropia relativa, utilizada nestes casos para estabelecer a identidade de uma locução, esta pode ser visto como um indicador de distância entre modelos. Sendo assim, ao comparar um modelo, construído sobre uma locução que esteja sendo testada, com modelos de locutores conhecidos, o menor valor encontrado apontará para o mais provável autor. Em cada teste, o menor valor de entropia apontando uma falsa identificação foi utilizado como limitante linear, levando a aceitação de somente resultados com entropias relativas inferiores.

Como os índices de identificações totais e com limitantes lineares foi equivalente em muitos dos testes, adotou-se um terceiro critério, a chamada media de confiabilidade de Bayes [7], calculada pela eq. (5).

$$P(C_T|\, x) = \frac{P(C_T)\, f_T(x)}{P(C_F)\, f_F(x) + P(C_T)\, f_T(x)}. \qquad (5)$$

$P(C_T|\, x)$ = probabilidade de identificação correta dado o vetor x.
$P(C_T)$ = probabilidade de identificação correta - obtida por contagem
$P(C_F)$ = probabilidade de identificação incorreta - obtida por contagem
$f_T(x)$ = $f(x|\, C_T)$: valor da distribuição (normal) para um locutor x, sendo este identificado corretamente
$f_F(x)$ = $f(x|C_F)$: valor da distribuição (normal) para um locutor x, sendo este identificados incorretamente.

5.2 Modelos de misturas Gaussianas

Como no modelo anterior, os parâmetros de entrada foram compostos por 20 coeficientes mel-cepstrais e 20 coeficientes delta-mel-cepstrais.

Após testes preliminares com 4, 16 e 64, o número de misturas gaussianas a compor cada modelo foi fixado em 32, escolha semelhante a de outras implementações vistas na literatura [10].

Outros testes foram conduzidos, com agrupamento inicial de vetores dentro das misturas feito por meio de redes SOM com 32 neurônios na camada de saída. Mas, uma vez que os resultados se assemelhavam aos obtidos

com agrupamento randômico, decidiu-se pelo último, dada a baixa complexidade computacional, bem como pelo existência de passo subsequente de otimização (ver figura 1).

O fim das iterações do algoritmo *estimar-maximizar* acontecia assim que não havia mais qualquer alteração na composição das classes, ou seja, quando mantidos inalterados os conjuntos de parâmetros que compunham cada classe.

Na eq. (6) os escores obtidos ao testar locuções serão valores que dependem dos modelos *j* utilizados para calculá-los .

Cada modelo estará representando conjuntos de diferentes valores, parametrizado com medidas diversas. Portanto, recomenda-se que seja feita uma normalização dos valores de *p* da eq. 6, pois estes serão calculados para cada locutor de referência, utilizando para isto um conjunto de locutores em separado [7]. Esta normalização foi feita pela divisão do escore obtido junto aos modelos de referência, pelo maior escore obtido com os locutores de normalização.

$$normalização\ (p_j(L)) = \frac{p_j(L)}{max\ p_k(L)} \qquad (6)$$

L : locução

p_j *:probabilidade para o modelo de referência j*

p_k*: probabilidade para o modelo de normalização k*

Sendo assim, o conjunto de modelos de referência, contendo os modelos *j* da eq. (6), foi formado a partir dos dados dos locutores de número 1 a 11. O conjunto de modelos de normalização *k* foi formado pelas amostras dos locutores de número 12, 13 e 14.

5.3 Comentários sobre os resultados

Os resultados dos testes com redes SOM foram consolidados na tabela 2. Observando as identificações totais e com limitantes, os testes de número 6 e 7 apresentam os mesmos valores. Mas, adicionando a medida de confiabilidade, sobressai o sétimo teste.

Comentando sobre a composição de parâmetros, primeiro deve-se atentar para o significado dos variados números de neurônios nas camadas de saída destas redes SOM. A curva dos resultados mostra que 56 classes, um valor próximo à quantidade de fones, é pouco. No outro extremo, distribuir vetores de parâmetros entre 400 classes pode estar causando uma dispersão indesejável, o que ocasiona a queda de performance. O valor intermediário, de 225 classes representadas por neurônios dispostos em uma matriz de dimensão 15x15, capturou as variações entre os locutores da melhor maneira.

Tabela 2. Valores de Resultados com Redes SOM

Rede SOM Nro. de Classes	*Identificações corretas*		*Medida de Confiabilidade*
	Total	*Limitante linear*	
1) 56	4	3	0,276
2) 56	5	2	0,283
3) 56	5	2	0,285
4) 56	5	2	0,285
5) 225	6	6	0,301
6) 225	7	6	0,304
7) 225	7	6	0,311
8) 400	6	5	0,288
9) 400	5	4	0,281

Tabela 3. Valores de Resultados com Modelos de Misturas Gaussianas

Testes com Misturas Gaussianas	*Total de Identificações*	*Medida de Confiabilidade*
10) 32 misturas	7	0,635
11) 32 misturas	7	0,611
12) 32 misturas	7	0,617

Para testes, foram sempre utilizadas frases e repetições de diferentes datas das empregadas para treinamento. O volume de dados utilizados para treinamento e testes influencia diretamente a eficiência com que os modelos são capazes de representar um locutor [11]. Variações momentâneas produzidas em diferentes sessões, somente poderão ter impacto reduzido se compensadas por parâmetros regulares.

6. Conclusão e Recomendações

Após analisados os resultados obtidos com os dois sistemas estudados, pode-se concluir que os modelos que empregaram redes neurais de mapas auto-organizáveis (SOM) e misturas gaussianas, são úteis ao reconhecimento de locutores da língua portuguesa. Os experimentos foram capazes de apontar um progresso, culminando por demostrar a eficácia dos modelos face os objetivos do estudo.

Refletindo o estado-da-arte no pré-processamento de locuções, o emprego de coeficientes mel-cepstrais se mostrou adequado.

Os sistemas que empregaram abordagem mista, com o uso de redes SOM, exercendo sua função no agrupamento (*clustering*) de parâmetros, em conjunto com cálculo de entropia entre classes, demonstraram precisão razoável.

A variação no número de neurônios na camada de saída da rede SOM esclareceu as dúvidas quanto a quantidade ideal para realizar uma tarefa desta natureza. Com 56 neurônios, quantia próxima ao conjunto de fones descritos pela fonética como suficientes para caracterizar as variações sonoras de uma língua, a classificação que a rede faz pode estar agrupando sons com peculiaridades individuais como uma coisa só, sem muita precisão para diferenciar pessoas. Quando foram utilizados 400 neurônios, pode ter ocorrido justamente o inverso, com a dispersão de sons muito próximos entre diferentes classes, impedindo a caracterização individual mais precisa. O que daí se depreende só tem valor após observados os resultados superiores que as redes com 225 classes apresentaram, com a identificação correta de 50 % dos locutores.

Vistos os experimentos, é igualmente correto afirmar a superioridade da implementação puramente estocástica, empregando os modelos de misturas gaussianas, se comparada aos outros modelos. Obteve-se 64 % de identificações corretas, em todos os testes, e a medida de confiabilidade de Bayes apontando com precisão este índice. Porém, esta superioridade vem junto com um maior custo computacional, uma vez que o cálculo de probabilidade em cada porção de uma locução desconhecida, deve ser repetido para cada modelo de referência, em cada uma de suas densidades. Se, por exemplo, tivermos uma locução com 1000 vetores a testar, 10 possíveis locutores, cada modelo de locutor composto por 32 misturas, serão realizados 1000 x 10 x 32 = 32×10^4 cálculos.

Merece que se discorra sobre o uso da medida de confiabilidade de Bayes. Com este indicador, calculado a partir da hipótese de que os resultados tem distribuição normal, que é a escolha natural quando se desconhece como possa ser a distribuição correta, atesta-se, em última análise, o quanto os valores resultantes estão próximos de um valor médio.

Ao empregar modelagem puramente estocástica, com cálculos de índices de semelhança que empregaram funções de distribuições normais, as medidas de confiabilidade indicaram com exatidão o que se podia esperar de cada teste. Já com as redes SOM, as medidas mostraram valores bem abaixo do obtido, como por exemplo confiabilidade de 0,311 para 50 % de identificações corretas. Mesmo com esta divergência, os resultados continuaram a ser analisados com este indicador, mas desta vez como critério de desempate apenas para os testes com redes SOM, e não para comparar estes últimos aos testes realizados com misturas gaussianas.

Considerando os objetivos aqui traçados, as amostras colhidas com frases expressas na língua portuguesa, tem maior valor que alternativas com textos de outras línguas. Para obter resultados mais significativos, as implementações devem ser submetidas a testes que contemplem maiores números de locutores fornecendo amostras. Além de aumentar a confiabilidade destes modelos, com uma quantidade maior de pessoas pode-se confirmar, principalmente, se a conformação mais apropriada de rede SOM deve ter 225 neurônios na camada de saída

Na prática das perícias criminalísticas, ambiente central para a aplicação do que aqui foi desenvolvido, o desemparelhamento das condições de amostragem é uma constante, assim como a presença de ruídos. Sendo assim, outras amostras devem ser colhidas em outros ambientes, a partir de telefones dos mais variados, utilizando outras aparelhagens de gravação, como gravadores de microcassetes, gravadores de mini-disk, alternando microfones, ou seja, condições variadas que incorporem efeitos proporcionados pelos diversos meios.

Pelo exposto, resta bem encaminhado o objetivo de aquisição de conhecimento para a realização de exames periciais, bem com um delineamento do que pode e deve ser trabalhado posteriormente, para a produção de sistemas mais robustos.

Referências

[1] Timoszczuk, Antonio P. "Reconhecimento Automático do Locutor com Redes Neurais Artificiais do Tipo Radial Basis Function (RBF) e Minimal Temporal Information (MTI)" Dissertação de Mestrado - Escola Politécnica da Universidade de São Paulo - 1998

[2] Furui, S. "Cepstral Analysis Technique for Automatic Speaker Verification" IEEE Transactions on Acoustic, Speech and Signal Processing, v.29, n.2, abril de 1981

[3] Deller, J.R.;Proakis, J.G.;Hansen, J.H.L "Discrete-Time Processing of Speech Signals" Macmillan-EUA - 1993

[4] Titze, I. "Principles of Voice Production" Prentice Hall - EUA- 1994

[5] Ramachandran, R.P.;Zilovic, M.S.;Mammone, R.J. "A Comparative Study of Robust Linear Predictive Analysis Methods with Applications to Speaker Identification" IEEE Transactions on Speech and Audio Processing, v.3, n.2, março de 1995

[6] Parsons, T.W. "Voice and Speech Processing" McGraw-Hill - EUA- 1987

[7] Gish, H.; Shmidt, M. "Text-Independent Speaker Identification" IEEE Signal Processing, outubro de 1994

[8] Campbell Jr., J.P. "Speaker Recognition: A Tutorial" IEEE Proceedings, v.85, n.9 - 1997

[9] Rabiner, L. "A Tutorial on HMM and Selected Applications in Speech Recognition" Proceedings IEEE, v.77, n.2, fevereiro de 1989

[10] Vuuren, S.V. "Comparison of Text-Independent Speaker Recognition Methods on Telephone Speech with Acoustic Mismatch" Proceedings of The International Conference on Spoken Language Processing - 1996

[11] Monte, E.;Hernando, J.;Miró, X.;Adolf, A. "Text-Independent Speaker Identification on Noisy Environment by Means of SOM" Proceedings of The International Conference on Spoken Language Processing - 1996