# NORMALIZAÇÃO DE LOCUTOR EM SISTEMA DE RECONHECIMENTO DE FALA

*Raquel de S.F. Dias, Fábio Violaro e Carlos Alberto Ynoguti*
*(*raquel@asga.com.br, fabio@decom.fee.unicamp.br, *ynogut@inatel.br)*

Departamento de Comunicações
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas – UNICAMP
Caixa Postal 6101, 13083-970, Campinas, SP

## RESUMO

Este artigo tem por objetivo avaliar um sistema de reconhecimento de fala de vocabulário flexível, quando utilizada a técnica de normalização de locutor. A técnica de normalização adotada foi a de escalonamento ("warping") do eixo de freqüências. Este escalonamento foi realizado pela variação do banco de filtros, na escala Mel, na obtenção dos coeficientes Mel Cepstrais. Estes coeficientes e suas derivadas foram empregados nos Modelos Ocultos de Markov (HMMs) que modelam as sub-unidades da fala (fones). Essas sub-unidades, após concatenadas, resultam os modelos das palavras constituintes do vocabulário de reconhecimento. Com a utilização deste método conseguiu-se reduzir a taxa de erros de um sistema básico, operando com um vocabulário de 400 palavras, de 19,25% para 11,25%.

## 1. INTRODUÇÃO

A Normalização de Comprimento do Trato Vocal ou ainda Normalização de Locutor [1,2,3,4], tem por objetivo tentar normalizar as representações paramétricas do sinal, de modo a reduzir os efeitos causados pela variabilidade da fala entre diferentes locutores.

O trato vocal possui diferentes formas e comprimentos para cada pessoa, resultando locuções com diferentes características acústicas. Na tentativa de minimizar esta variabilidade entre os locutores, uma das principais responsáveis pela degradação de desempenho dos sistemas de reconhecimento de fala, será analisado ao longo deste artigo o processo de normalização de locutor.

## 2. NORMALIZAÇÃO DE LOCUTOR

O processo de normalização de locutor é representado pela transformação dos parâmetros acústicos da fala. Esta transformação é realizada por funções de distorção aplicadas ao banco de filtros na escala Mel [5].

O banco de filtros na escala Mel consiste numa série de filtros passa-faixa triangulares centrados em freqüências que variam linearmente até 1 kHz e, a partir daí, crescem exponencialmente com um fator de $2^{1/5}$. Nas simulações efetuadas foi empregada uma freqüência de amostragem de 11,025 kHz (fs/2 = 5,512 kHz), impondo-se assim a utilização de 21 filtros.

### 2.1. Transformação dos Parâmetros

O processo de transformação dos parâmetros acústicos, ou ainda de normalização do trato vocal, é obtido pelo escalonamento do eixo de freqüências do banco de filtros. O escalonamento destas freqüências é realizado linearmente, por um fator de distorção $\alpha$ compreendido entre 0,88 e 1,12 (variando em passos de 0,02). No final deste processo tem-se um novo banco de filtros, conforme equação abaixo, com freqüências escalonadas. Dependendo do fator de escalonamento utilizado, estas freqüências ora serão expandidas ($\alpha < 1$), ora serão comprimidas ($\alpha > 1$):

$$f' = \beta . f$$

onde
f – representa a freqüência original na escala Mel,
$f'$ - representa a freqüência escalonada,
$\alpha$ - representa o fator de distorção,
$\beta = 1/\alpha$ - representa o fator de escalonamento em freqüência (para $\alpha$ variando entre 0,88 e 1,12).

### 2.2 Considerações

Quando o eixo de freqüências é escalonado, a largura de faixa do sinal resultante difere da largura de faixa do sinal original, provocando uma alteração na informação útil que será utilizada pelo sistema Uma solução para este problema, utilizada por [1] na tentativa de suavizar a variação entre a largura de faixa do sinal original e do sinal normalizado, é considerar uma função de escalonamento não linear, de tal forma que a largura de faixa do sinal escalonado seja a mais próxima do sinal original.

Como exemplo da função linear utilizada, tem-se:

$$G(f) = \begin{cases} \beta . f & 0 \le f \le f_0 \\ \dfrac{f_{max} - \beta . f_0}{f_{max} - f_0} ( f - f_0 ) + \beta . f_0 & f_0 \le f \le f_{máx} \end{cases}$$

onde:
$f_0$–representa uma freqüência escolhida empiricamente, de valor acima da mais alta formante significativa .
$f_{máx}$–representa a máxima largura de faixa considerada no sinal original.

Neste trabalho, o valor de $f_0$ foi escolhido como sendo a máxima freqüência central utilizada ($f_0$ = 4,595 kHz). O valor de $f_{máx}$ escolhido corresponde à máxima freqüência abrangida pelo último filtro (21º) do banco de filtros empregado ($f_{máx}$ =5,278kHz).

Na Figura 1.a tem-se a representação do banco de filtros de referência, sem escalonamento ($\beta$=1), e nas Figuras 1.b e 1.c, o mesmo banco com compressão máxima ($\beta$=1,136) e expansão máxima ($\beta$=0,893) respectivamente.

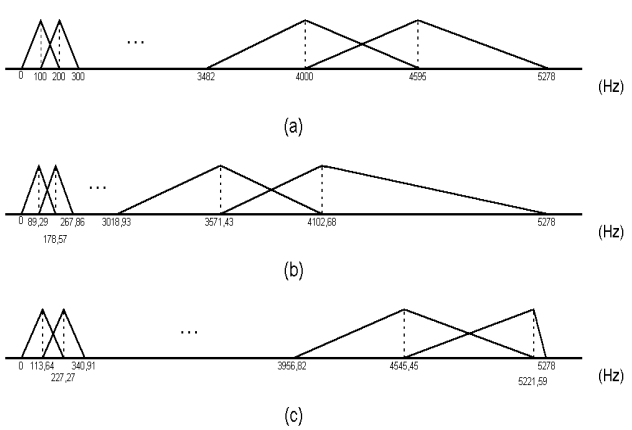


**Figura 1:** Figura representativa do banco de filtros referência (a), banco de filtros comprimido (b) e banco de filtros expandido (c).

# 3. Avaliação de HMM Utilizando Normalização de Locutor

A normalização de locutor é dividida em duas etapas, aplicadas iterativamente: escolha do fator de distorção $\alpha$ que melhor caracterize cada locutor analisado e retreinamento dos HMMs utilizando o $\alpha_{ótimo}$ obtido para cada locutor.

É importante ressaltar que, antes da escolha do melhor $\alpha$ e do retreinamento dos HMMs, as locuções de treinamento devem ser separadas por locutor. Cada locutor será representado por um conjunto de locuções (frases ou palavras isoladas).

## 3.1 Escolha do melhor a

Uma vez definidos os locutores e suas respectivas locuções, é feita a escolha do melhor $\alpha$, para cada locutor. O melhor $\alpha$ será aquele que, após o escalonamento do banco de filtros empregado na obtenção dos parâmetros de análise (mel cepstrais e sua derivadas), proporcionará a maior verossimilhança média nas suas locuções.

Na escolha do melhor fator de distorção são levados em consideração todas as locuções do locutor sendo analisado e o modelo HMM utilizado como referência. O algoritmo utilizado no cálculo do melhor $\alpha$ é apresentado a seguir:

**Inicialização:**

As matrizes de transição e emissão, utilizadas no modelamento dos HMMs, são inicializadas com os valores obtidos no primeiro treinamento do sistema, sem normalização ($\alpha$=1). O HMM obtido deste primeiro treinamento será chamado de HMM pré-treinado.

**Recursão:**

Inicialmente são definidas as seguintes variáveis:

$\alpha$ - fator de distorção ( $0,88 \leq \alpha \leq 1,12$).

$W_i$ – conjunto de transcrições referentes ao locutor i.

$\lambda$ - modelo HMM pré-treinado.

$X_i^{\alpha}$ - conjunto de parâmetros obtidos após escalonamento do banco de filtros por um fator $\alpha$, para o locutor i.

max $[P(X_i^{\alpha} / \lambda, W_i)]$ – representa a máxima probabilidade de se obter um conjunto de observação $X_i$, escalonado de $\alpha$, dado um modelo $\lambda$ e um conjunto de transcrições $W_i$.

Para cada locutor i, faz-se:

- Variar $\alpha$ em intervalos de 0,02.
- Calcular a max [ $P(X_i^{\alpha} / \lambda, W_i)$ ] entre os 13 valores de $\alpha$.
- Armazenar o valor de $\alpha$ que proporcionou a max $[P(X_i^{\alpha} / \lambda, W_i)]$, associando-o a seu respectivo locutor.

O algoritmo utilizado no cálculo da máxima verossimilhança foi o algoritmo de Viterbi.

**Término:**

O processo é finalizado quando se obtém, para cada locutor i, o seu respectivo fator de distorção $\alpha^i$.

## 3.2 – Treinamento

O processo de treinamento dos modelos HMMs utilizando normalização de locutor é semelhante ao treinamento usual dos HMMs. Entretanto, é diferenciado na sua inicialização (realizada a partir do HMM pré-treinado) e na normalização das características espectrais dos locutores utilizados.

Inicialmente calculam-se os $\alpha$'s ótimos para cada locutor do conjunto de M locutores de treinamento. Em seguida estes $\alpha$'s, associados a seus respectivos locutores, são empregados para calcular um novo conjunto de parâmetros mel-cepstrais que serão utilizados no retreinamento do sistema. O retreinamento é então executado durante tantas épocas quantas foram necessárias para que a distorção desejada seja atingida (0,001 de diferença relativa entre a época anterior e a época atual).

Depois de realizadas todas as épocas de retreinamento, tem-se um novo modelo HMM ($HMM_{NOVO}$). Este novo HMM será utilizado na escolha de novos $\alpha$'s para cada locutor ($\alpha_{NOVO}$). Em seguida é feito uma comparação entre os valores de $\alpha_{NOVO}$ e os valores de $\alpha$ inicialmente calculados. Caso pelo menos um dos $\alpha$'s, para um mesmo locutor, seja diferente, faz-se a atualização destes $\alpha$'s ($\alpha \Leftarrow \alpha_{NOVO}$), para o cálculo dos novos coeficientes mel-cepstrais, e do HMM ($HMM \Leftarrow HMM_{NOVO}$), para que em seguida seja executado um novo retreinamento do sistema. O sistema deverá continuar sendo retreinado até que o fator de distorção de cada locutor ($\alpha^i$) não seja mais alterado entre um retreinamento e outro. Obtém-se no final deste processo o $HMM_{NORMALIZADO}$.

## 3.3 Reconhecimento

O reconhecimento das locuções é realizado da mesma forma que no sistema sem normalização de locutor. Entretanto, antes de se reconhecer cada locução, deve-se escolher o melhor fator de

distorção para cada um dos locutores de teste (locutores utilizados no reconhecimento).

Deve-se mencionar ainda que, tanto no treinamento quanto no reconhecimento, uma vez obtido o melhor fator de distorção, para cada locutor, este fator é utilizado para todas as demais locuções do respectivo locutor, na obtenção dos parâmetros acústicos de cada locução.

# 4. AVALIAÇÃO DO SISTEMA

Uma vez definida a melhor forma de escolha do $\alpha$, foi iniciada a avaliação do sistema. Para essa avaliação foi empregada a mesma base de fala empregada em [6] e composta por frases foneticamente balanceadas e seqüências de dígitos conectados (42 locutores de treinamento e 2000 frases). Já os testes foram efetuados com uma outra base de dados de palavras isoladas por nós gerada e consistindo de nomes de pessoas e apelidos (15 locutores de teste e 300 nomes). Os resultados estão mostrados na Tabela 1, onde os HMMs foram implementados com misturas de 3 gaussianas por parâmetro. A escolha do $\alpha_{ótimo}$ de cada locutor foi feita a partir de 4 locuções e deve-se lembrar que a cada novo retreinamento é escolhido um novo $\alpha$ para cada locutor.

| **Número de Ret.** | | **Distorção (convergência)** | **a's Mod.** | **Taxa de Erros % (Teste)** |
|---|---|---|---|---|
| Trein.($\alpha$ =1) | | 0,0009 | --- | 17,00 |
| 1º retrein. | 1º época | 0,0089 | --- | 16,67 |
| | 2º época | 0,0013 | | 16,00 |
| | 3º época | 0,0006 | | 15,00 |
| 2º retrein. | 1º época | 0,0021 | 22 | 15,00 |
| | 2º época | 0,0006 | | 15,33 |

**Tabela 1:** Resultado obtido para o sistema de reconhecimento com normalização do locutor, utilizando-se misturas de 3 gaussianas

Como pode-se observar na Tabela 1, obteve-se uma melhora de desempenho quando comparado ao sistema sem normalização (17 % de erro), sendo necessário apenas 1 retreinamento antes que o desempenho do sistema fosse prejudicado. No entanto, o objetivo principal do sistema não foi atingido, isto é, o sistema deveria continuar sendo retreinado até que não houvesse mais modificação entre os $\alpha$'s de um mesmo locutor entre um retreinamento e outro.

A partir dos resultados obtidos, surgiu a hipótese de que provavelmente os HMMs ainda não estariam treinados o suficiente para que pudessem passar para um próximo retreinamento. Seguindo esta idéia resolveu-se aumentar o número de épocas de cada retreinamento, até que fossem atingidas, mais ou menos, 5 ou 6 épocas, ou enquanto houvesse melhora de desempenho do sistema. Assim resultaram os resultados da Tabela 2.

| | | **Distorção (converg.)** | **a's Mod.** | **Taxa de Erros % (teste)** |
|---|---|---|---|---|
| Treinamento ($\alpha$ =1) | | 0,0009 | --- | 17,00 |
| 1º ret. | 1º época | 0,0089 | --- | 16,67 |
| | 2º época | 0,0013 | | 16,00 |
| | 3º época | 0,0006 | | 15,00 |
| | 4º época | 0,0004 | | 15,00 |
| | 5º época | 0,0004 | | 15,00 |
| 2º ret. | 1º época | 0,0019 | 22 | 14,00 |
| | 2º época | 0,0005 | | 14,00 |
| | 3º época | 0,0003 | | . |
| | 4º época | 0,0003 | | . |
| | 5º época | 0,0002 | | . |
| | 6º época | 0,0002 | | |
| | 7º época | 0,0002 | | |
| | 8º época | 0,0001 | | 13,00 |
| 3º ret. | 1º época | 0,0005 | 10 | 13,33 |
| | 2º época | 0,0002 | | 13,33 |

**Tabela 2:** Tabela representativa do desempenho do sistema normalizado, utilizando-se misturas de 3 gaussianas, aumentando-se o número de épocas em cada retreinamento.

Comparando-se as Tabelas 1 e 2, pode-se notar que, ao se aumentar o número de épocas em cada retreinamento, obtém-se uma melhora de desempenho em relação ao procedimento anterior. Entretanto, como pode ser visto na Tabela 2, não se pode afirmar exatamente quantas épocas devem ser realizadas para que haja melhora de desempenho no sistema. O único fato que pode ser constatado é que, com a normalização dos locutores, a faixa de distorção é reduzida de 0,001 para 0,0001.

Apesar de ter-se uma melhora de desempenho utilizando-se 2 retreinamentos, ainda há o fato de que os locutores continuam com valor de $\alpha$ variável entre um retreinamento e outro. Assim, tentando resolver este problema e o fato de não se saber exatamente quantas épocas serão necessárias para cada retreinamento, se decidiu seguir 2 critérios de parada para o retreinamento do sistema. A primeira idéia é a de se retreinar os modelos HMMs com apenas 1 época por retreinamento, realizando-se uma nova escolha de $\alpha$ a cada novo retreinamento. Vários retreinamentos serão executados até que não haja mais variação de $\alpha$, de um mesmo locutor, entre um retreinamento e outro. A segunda idéia é a de que, uma vez que os $\alpha$'s permaneçam constantes, para todos os locutores, seja verificada a distorção relativa obtida neste retreinamento. Caso esta distorção seja maior que 0,0001, entre o retreinamento atual e retreinamento anterior, o retreinamento continuará sendo realizado até que se obtenha a distorção desejada.

Tomando como base ambos os procedimento descritos anteriormente, foram obtidos os resultados ilustrados na Tabela 3.

| | Distorção | a's Modificados | Taxa de Erros % |
|---|---|---|---|
| Treinamento | 0,0009 | --- | 17,00 |
| 1º retreinamento | 0,0089 | --- | 16,67 |
| 2º retreinamento | 0,0032 | 21 | 16,00 |
| 3º retreinamento | 0,0013 | 8 | 16,00 |
| 4º retreinamento | 0,0007 | 3 | 15,34 |
| 5º retreinamento | 0,0005 | 3 | 15,00 |
| 6º retreinamento | 0,0005 | 1 | 14,33 |
| 7º retreinamento | 0,0004 | 1 | 14,33 |
| 8º retreinamento | 0,0003 | 0 | 14,00 |
| 9º retreinamento | 0,0002 | 0 | 14,33 |
| 10º retreinamento | 0,0002 | 0 | 14,00 |
| 11º retreinamento | 0,0002 | 0 | 13,67 |
| 12º retreinamento | 0,0001 | 0 | 13,67 |

**Tabela 3:** Tabela ilustrativa do desempenho do sistema ao utilizarmos 1 época por retreinamento. A cada novo retreinamento é calculado um novo valor de $\alpha$.

Como pode ser visualizado na Tabela 3, as duas condições adotadas foram bastante relevantes na escolha da melhor forma de normalização dos locutores. O limite de treinamento (número de épocas necessárias para cada retreinamento) é obtido de forma mais coerente que o procedimento anterior, obtendo-se ainda uma melhora de desempenho a cada retreinamento. Deve-se mencionar ainda que, com estas novas condições, diminuiu-se o tempo de treinamento do sistema. Onde antes eram utilizadas 13 épocas para normalizar o sistema, agora são utilizadas apenas 12 épocas.

## 5. RESULTADOS FINAIS

Após ser definida a melhor forma de retreinamento, reconhecimento e, principalmente, de escolha do melhor fator de distorção para normalização do sistema de reconhecimento de fala, serão apresentados os resultados finais deste processo ao se utilizar misturas de 5 gaussianas por parâmetro.

Embora em testes anteriores, sem normalização de locutor, os HMMs tenham apresentado um melhor desempenho quando utilizando misturas de 6 gaussianas, no decorrer dos testes pôde-se perceber que os locutores iam tendo suas características espectrais aproximadas (normalizadas), diminuindo assim a quantidade de gaussianas necessárias para sua representação. Desta forma, para melhor visualização do desempenho do sistema e para evitar que alguma gaussiana deixasse de ser corretamente modelada, devido à pequena quantidade de dados de treinamento utilizada, optamos por usar apenas 5 gaussianas na representação dos resultados finais do sistema.

Além das 5 gaussianas, utilizou-se também 20 locutores de teste, 5 a mais que os usados nos testes anteriores. A base de teste também foi ampliada de 300 para 400 nomes ou apelidos. Nesse caso a taxa de acertos do sistema básico (sem normalização) caiu para 19,25%. Desta forma, tem-se um sistema com as seguintes especificações:

- Locutores de treinamento: 42 locutores, pronunciando um
- total de 2000 locuções.
- Parâmetros utilizados: mel, dmel e ddmel ( 12 coeficientes cada).
- Sub-unidades fonéticas utilizadas: fones independentes do contexto.
- Tipo de HMM: Contínuo.
- Número de gaussianas por estado e por parâmetro (densidades independentes): 5.
- Algoritmo de treinamento: Baum-Welch.
- Número de locuções utilizadas na escolha do melhor $\alpha$: 4 frases para o treinamento e 4 nomes para o reconhecimento. Critério de parada para o treinamento: distorção relativa = 0,0001.
- Locutores de teste: 20 (10 homens e 10 mulheres) pronunciando um total de 400 nomes.
- Algoritmo de reconhecimento: One-Step.
- Avaliação do sistema:

O resultados dos testes efetuados encontra-se ilustrado na Tabela 4, onde o tempo apresentado do 9º ao 19º retreinamento, representa o valor total obtido durante 11 retreinamentos consecutivos, uma vez que não há variação do fator de distorção entre estes retreinamentos. O tempo de treinamento dos modelos foi verificado utilizando-se um Pentium II – 300 MHz.

| | Distorção | a's Mod. | Taxa de Erros % | Tempo de Treinamento |
|---|---|---|---|---|
| Treinamento | 0,001 | --- | 19,25 | 27:58:51 |
| 1º retrein. | 0,0056 | --- | 19,00 | 08:19:54 |
| 2º retrein. | 0,0035 | 26 | 17,50 | 08:32:24 |
| 3º retrein. | 0,0021 | 20 | 15,50 | 08:27:35 |
| 4º retrein. | 0,0012 | 9 | 15,00 | 07:47:45 |
| 5º retrein. | 0,0008 | 8 | 16,00 | 04:31:11 |
| 6º retrein. | 0,0007 | 5 | 15,75 | 08:59:24 |
| 7º retrein. | 0,0005 | 2 | 15,50 | 04:48:58 |
| 8º retrein. | 0,0004 | 2 | 14,00 | 07:49:08 |
| 9º retrein. | 0,0004 | 0 | . . . | 59:34:50 |
| . . . | . . . | | | |
| 19º retrein. | 0,0001 | | 11,25 | |

**Tabela 4:** Tabela ilustrativa do desempenho do sistema normalizado, utilizando 5 gaussianas. A duração de cada retreinamaneto leva em consideração o período de tempo utilizado para a escolha do $\alpha$.

Deve-se mencionar que o procedimento de escolha do $\alpha_{ÓTIMO}$ pode também ser considerado um importante critério para a caracterização dos locutores em masculino e feminino, como ilustrado nas Figuras 2.a e 2.b.

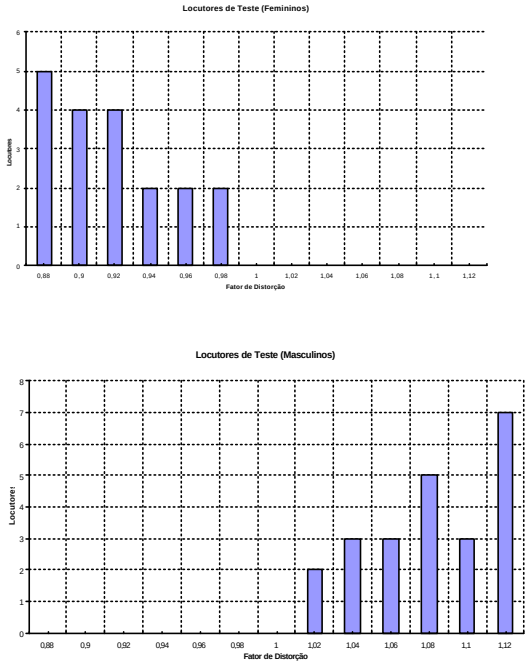


**Figura 2:** Histogramas representativos da faixa de valores de $\alpha$ escolhida para os locutores femininos (a) e masculinos (b), para o treinamento do sistema utilizando-se mistura de 5 gaussianas

Nas Tabelas 5 e 6 tem-se os valores de $\alpha$ definidos para os locutores de teste em cada novo retreinamento. Estes $\alpha$'s são obtidos utilizando o $HMM_{normalizado}$ para 5 misturas de gaussianas, conforme procedimento descrito na seção 3.

| ***Locutor*** | **Número de Retreinamentos** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **...** | **19** |
| **F01** | 0,88 | 0,88 | 0,90 | 0,90 | 0,92 | 0,92 | 0,92 | **...** | 0,92 |
| **F02** | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | **...** | 0,88 |
| **F03** | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | **...** | 0,88 |
| **F04** | 0,96 | 0,92 | 0,92 | 0,92 | 0,92 | 0,92 | 0,92 | **...** | 0,92 |
| **F05** | 0,92 | 0,92 | 0,90 | 0,90 | 0,92 | 0,92 | 0,92 | **...** | 0,92 |
| **F06** | 0,92 | 0,90 | 0,90 | 0,88 | 0,88 | 0,88 | 0,88 | **...** | 0,88 |
| **F07** | 0,96 | 0,90 | 0,90 | 0,90 | 0,88 | 0,88 | 0,88 | **...** | 0,88 |
| **F08** | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | 0,88 | **...** | 0,88 |
| **F09** | 0,88 | 0,94 | 0,94 | 0,92 | 0,88 | 0,88 | 0,88 | **...** | 0,88 |
| **F10** | 0,92 | 0,90 | 0,90 | 0,90 | 0,92 | 0,92 | 0,92 | **...** | 0,92 |

**Tabela 5:** Tabela representativa dos valores de $\alpha$ obtidos para os locutores de teste femininos (HMM normalizado com mistura de 5 gaussianas por estado)

| ***Locutor*** | **Número de Retreinamentos** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **...** | **19** |
| **M02** | 1,04 | 1,06 | 1,06 | 1,06 | 1,08 | 1,08 | 1,08 | **...** | 1,08 |
| **M03** | 1,06 | 1,06 | 1,08 | 1,08 | 1,08 | 1,08 | 1,08 | **...** | 1,08 |
| **M04** | 0,96 | 0,98 | 0,98 | 0,98 | 1,00 | 1,00 | 1,00 | **...** | 1,00 |
| **M05** | 1,08 | 1,06 | 1,08 | 1,08 | 1,08 | 1,08 | 1,08 | **...** | 1,08 |
| **M06** | 1,04 | 1,02 | 1,02 | 1,02 | 1,04 | 1,04 | 1,04 | **...** | 1,04 |
| **M07** | 0,96 | 1,0 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | **...** | 1,00 |
| **M08** | 0,96 | 0,96 | 0,98 | 0,98 | 0,96 | 0,96 | 0,96 | **...** | 0,96 |
| **M09** | 0,96 | 0,94 | 0,94 | 0,96 | 0,96 | 0,96 | 0,96 | **...** | 0,96 |
| **M10** | 1,04 | 1,04 | 1,04 | 1,04 | 1,04 | 1,04 | 1,04 | **...** | 1,04 |

**Tabela 6:** Tabela representativa dos valores de $\alpha$ obtidos para os locutores de teste masculinos (HMM normalizado com mistura de 5 gaussianas por estado)

Como pode ser observado nas Tabelas 5 e 6, depois de um certo número de retreinamentos os $\alpha$'s dos locutores de teste mantêm-se constantes, mas a verossimilhança média calculada com as locuções de treinamento continua aumentando. Uma justificativa para este fato é que o material utilizado no treinamento é bem maior que o material de teste.

# 6. CONCLUSÕES

A normalização de locutor não deve ser confundida com a adaptação de locutor. Na primeira técnica todos os locutores utilizados no retreinamento do sistema são normalizados, iterativamente, em relação a um locutor médio. Na segunda técnica, o retreinamento é realizado para um locutor em particular, aquele para o qual o sistema será adaptado, daí o tempo de retreinamento resultante ser muito menor do que na normalização de locutor. Após a normalização do sistema, entretanto, o tempo para um novo locutor calcular seu fator de normalização $\alpha$ e passar a utilizar o sistema de reconhecimento é muito pequeno (no nosso caso ele foi calculado com apenas 4 locuções)

No presente trabalho foi avaliada a técnica de normalização de comprimento do trato vocal entre diferentes locutores. Para avaliação desta técnica utilizou-se um *Sistema de Reconhecimento de Fala Independente do Locutor e de Vocabulário Flexível.*

A utilização de vocabulário flexível no sistema adotado foi de grande valia, principalmente por proporcionar uma maior flexibilidade quando da criação do vocabulário a ser reconhecido pelo sistema. Desta forma, pôde-se reconhecer locuções fora do universo com o qual o sistema foi treinado, daí a maior versatilidade do mesmo.

Quanto à técnica utilizada, pôde-se comprovar que a normalização de comprimento do trato vocal realizada pelo escalonamento do banco de filtros, na escala Mel, é uma importante ferramenta a ser engajada nos sistemas de reconhecimento de fala. Pode-se ressaltar que, além de

proporcionar uma melhora no desempenho do sistema, é de fácil implementação. Este tipo de normalização, segundo [1], tende a proporcionar melhores resultados do que as técnicas de *Separação das Características Acústicas (masculino e feminino)* e de *Normalização da Média Cepstral.*

Ao longo deste trabalho foi verificada a necessidade de uma estratégia que proporcionasse a melhor maneira de se normalizar o sistema, sendo escolhida uma estratégia de, iterativamente, calcular o $\alpha^i$ para cada locutor e retreinar o sistema a cada nova época.

Deve-se destacar ainda o curto período de tempo necessário para o cálculo do $\alpha_{ótimo}$ de cada locutor, na fase de reconhecimento. Empregando-se 4 locuções (nomes) por locutor, obteve-se, avaliando-se um conjunto de 20 locutores de teste, um tempo médio de 19 minutos para o cálculo do $\alpha_{ótimo}$ de todos os locutores. Assim, para cada locutor, gastou-se aproximadamente 1 minuto no cálculo do $\alpha_{ótimo}$ .

O procedimento de escolha do melhor fator de distorção $\alpha$ pode ser considerado um fator determinante no desempenho do sistema. Isto decorre do fato de que ao ser escolhido, para pelo menos um locutor, um valor de $\alpha$ diferente do seu "ideal", o treinamento para este locutor será realizado fora de suas características espectrais. Desta forma, os HMMs assumirão valores de verossimilhança bastante baixos, comprometendo assim o desempenho do sistema.

Como desvantagem da técnica de normalização tem-se a proximidade das gaussianas. A cada retreinamento do sistema, utilizando um novo conjunto de coeficientes $\alpha$, mais normalizado ele se torna. Esta normalização, por sua vez, faz com que algumas gaussianas utilizadas no modelamento dos HMMs deixem de ser relevantes, como eram no início do processo, passando a ter coeficientes de ponderação bastante baixos. Este fato poderá gerar problemas de underflow no sistema, caso não tenha sido previsto. Outra desvantagem da técnica é o tempo gasto em cada retreinamento, além do tempo necessário para o treinamento inicial do sistema (sem normalização). Em contrapartida a estes problemas tem-se, entretanto, o aumento considerável de desempenho do sistema.

A contribuição mais significativa deste trabalho, em relação a [1], foi a de proporcionar um método mais robusto para o retreinamento do sistema e para obtenção do $\alpha_{ótimo}$ de cada locutor. Um maior detalhamento do processo de normalização pode ser encontrado em [7].

# 7. BIBLIOGRAFIA


[1] Lee, L. and Rose, R. – "A Frequency Warping Approach to Speaker Normalization". *IEEE Transactions on Speech and Audio Processing*, Vol. 6, nº 1, January 1998, pp 49-60.

[2] Andreou, A., Kamm T. and Cohen, J. – "Experiments in Vocal Tract Normalization". *Proceedings CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[3] Hao, Y. and Fang, D. – "Speech Recognition using Speaker Adaptation by System Parameter Transformation". *IEEE Transactions on Speech and Audio Processing*, Vol. 2, nº 1, Part 1, January, 1994, pp 63-67.

[4] Padmanabhan M., Lalit R. B., Nahamoo, D. and Picheny, M. A. – "Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems". *IEEE Transactions on Speech and Audio Processing*, Vol. 6, nº 1, January 1998, pp 71-77.

[5] Davis, S. B. and Mermelstein, P. – "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-28, nº 4, August 1980, pp 357-368.

[6] Ynoguti, C. A – "Reconhecimento de Fala Contínua usando Modelos Ocultos de Markov". *Tese de Doutorado*, UNICAMP, Campinas, Maio de 1999, pp 24-32, 47-55, 58-82.

[7] Dias, R. S. F. – "Normalização de Locutor em Sistema de Reconhecimento de Fala". *Tese de Mestrado*, UNICAMP, Campinas, Novembro de 2000.