Treinamento em Múltiplas Condições com Ruídos de Espectro Colorido para Identificação Robusta de Locutor

L. Zão e R. Coelho

***Resumo*— Este trabalho propõe uma nova técnica de treinamento em múltiplas condições para reconhecimento automático de locutor (RAL) robusto a ruídos acústicos ambientais. Para o treinamento, é proposto um gerador de ruídos artificiais coloridos, que são utilizados para obter os modelos dos locutores. Estes ruídos são obtidos pela filtragem de uma sequência de amostras com espectro branco. A função de transferência adotada para o filtro é baseada na forma proposta por Al-Alaoui. A nova técnica para RAL robusto é avaliada para identificação de locutor, com locuções de teste submetidas a ruídos acústicos ambientais de diferentes fontes. Os resultados demonstram que a técnica proposta neste artigo supera o desempenho do treinamento em múltiplas condições que utiliza o ruído branco.**

***Palavras-Chave*— Reconhecimento automático de locutor, treinamento em múltiplas condições, ruídos coloridos.**

***Abstract*— This paper proposes a multicondition training technique for automatic speaker recognition in unknown noisy conditions. A colored-spectra noises generator is proposed for the speakers models training. The noises are obtained by filtering a random sequence with white power spectrum density. The Al-Alaoui's rule is adopted for the filter's transfer function. The multicondition training is evaluated for speaker identification tasks. The test utterances are corrupted with different environmental noises and signal-to-noise ratios. The results show that the proposed technique outperforms the multicondition training approach based on the use of white noise.**

***Keywords*— Automatic speaker recognition, multicondition training, colored-spectra noises**

I. INTRODUÇÃO

A crescente necessidade de sistemas de segurança tem impulsionado o uso da autenticação biométrica. Enquanto as soluções convencionais são geralmente baseadas no uso de senhas ou cartões de identificação, as biométricas [1] empregam o reconhecimento de padrões de características humanas, tais como a impressão digital, a íris, a face e a voz.

A voz é considerada uma das características biométricas mais naturais para reconhecer indivíduos. Além da mensagem transmitida, o sinal de voz contém ainda informações de identidade, sexo, idioma e as condições físico-emocionais do locutor. Além disso, o sinal de voz é de fácil aquisição e seu processamento é considerado simples para a tecnologia existente. Sistemas de reconhecimento automático de locutor (RAL) [2] têm ampla aceitação em aplicações na área de segurança e defesa, como o controle de acesso e investigações forenses. No entanto, a presença de ruídos acústicos no sinal de voz é considerada uma das principais causas da queda de desempenho em sistemas de reconhecimento de locutor. Este impacto é atribuído à variabilidade ou ao desconhecimento das diferentes fontes de ruídos acústicos. Uma técnica interessante, para prover robustez ao RAL em ambientes ruidosos, é o treinamento em múltiplas condições [3] [4]. A ideia principal é submeter as locuções de treinamento a diversas situações de ruídos de forma a diminuir o descasamento de condições entre as fases de treinamento e teste. Em [5], foi utilizado o ruído gaussiano branco, em diversas relações sinal-ruído (RSR), para corromper o sinal de voz. Os autores argumentam que o ruído branco foi escolhido devido ao desconhecimento das características dos ruídos presentes nos sinais de voz. Contudo, diversos estudos concluíram que os ruídos acústicos ambientais possuem espectros coloridos [6] [7].

Este trabalho propõe uma nova técnica de treinamento em múltiplas condições para prover robustez ao RAL. Para o treinamento, é proposto um método para geração de ruídos artificiais coloridos. Estes ruídos são utilizados para corromper as locuções limpas, e são obtidos a partir da filtragem de uma sequência de números aleatórios com espectro branco. As funções de transferência dos filtros utilizados na obtenção dos espectros coloridos são baseadas na forma proposta por Al-Alaoui [8]. Os modelos dos locutores são obtidos a partir destas locuções corrompidas.

A técnica de treinamento em múltiplas condições proposta neste trabalho é avaliada para a identificação de locutor. Para isto, são gerados ruídos com três cores de espectro distintas para corromper as locuções de treinamento: branco, marrom e rosa. Nos experimentos são também utilizados modelos obtidos com as locuções de treinamento limpas e a partir de sinais de voz corrompidos por ruído branco. Para as locuções de teste, foram considerados ruídos acústicos ambientais coletados de fontes reais distintas. Os resultados mostram que a técnica proposta neste artigo aumenta as taxas de acertos da identificação em relação aos métodos utilizados como referência.

O restante deste trabalho está organizado da seguinte forma. A Seção II descreve os atributos da voz e o classificador utilizados no RAL. A Seção III apresenta a proposta de treinamento em múltiplas condições baseada no uso de ruídos coloridos. Na mesma Seção, é descrito o método de geração dos ruídos coloridos artificiais. A Seção IV discute os principais resultados obtidos com os experimentos de identificação

Leonardo Zão, Programa de Pós-Graduação em Engenharia de Defesa (PGED), Instituto Militar de Engenharia (IME), Rio de Janeiro, Brasil, E-mail: zao@ime.eb.br. Rosângela Coelho, Programa de Pós-Graduação em Engenharia Elétrica, Instituto Militar de Engenharia (IME), Rio de Janeiro, Brasil, E-mail: coelho@ime.eb.br.

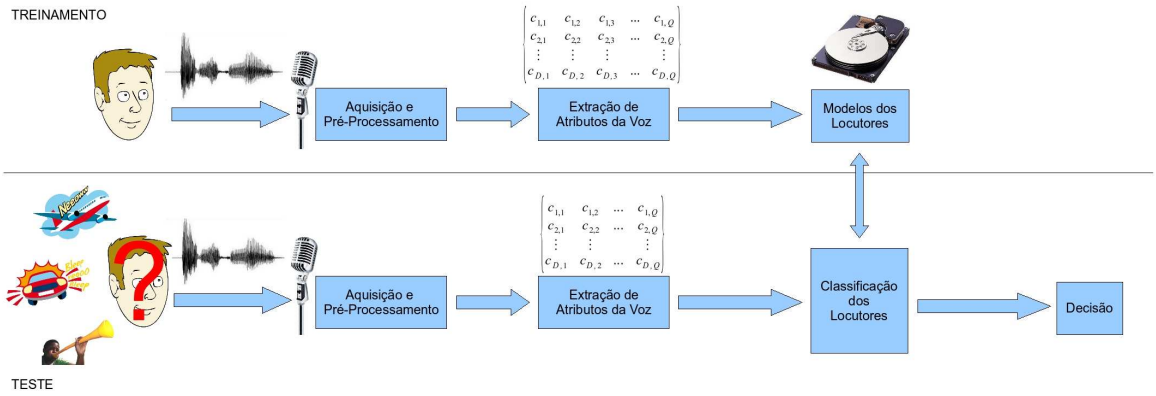

Fig. 1. As fases e etapas que compõem um sistema de reconhecimento automático de locutor.

de locutor, realizados para avaliação da técnica proposta. Finalmente, a Seção V conclui o presente trabalho.

II. O Reconhecimento Automático de Locutor

Um sistema de RAL é geralmente dividido em duas fases: treinamento e testes. Conforme ilustrado na Fig. 1, cada uma destas fases é composta de três etapas: aquisição/pré-processamento do sinal de voz, extração de atributos ou características da voz e classificação de locutor. A primeira etapa realiza a digitalização e o janelamento do sinal de voz em segmentos, ou quadros, de curta duração ($\approx$ 20 ms). Na segunda etapa, vetores de atributos são extraídos dos quadros obtidos na etapa anterior. Estes vetores são concatenados formando uma matriz de atributos. Durante a fase de treinamento, a etapa de classificação é responsável por obter e armazenar os modelos de locutores a partir das matrizes de atributos. Já na fase de teste, a matriz de atributos é confrontada com os modelos previamente armazenados.

Duas tarefas podem ser realizadas na etapa de classificação: identificação ou verificação de locutor. Na identificação de locutor, o sistema decide a qual dos usuários cadastrados pertence a locução de teste. Já na verificação, o locutor declara sua identidade e o sistema decide se a aceita, ou não, como verdadeira.

Na literatura, os coeficientes mel-cepestrais (MFCC - *mel-frequency cepstral coefficients*) [9] [10] e o modelo de misturas gaussianas (GMM - *gaussian mixture model*) [11] são considerados referência de bom desempenho em sistemas de RAL. Atributos dinâmicos (Δ) são geralmente utilizados em conjunto com os coeficientes MFCC.

A. Os Coeficientes MFCC

Após a aquisição e janelamento do sinal de voz, o mesmo é transformado para o domínio da frequência através da transformada rápida de Fourier (FFT - *fast Fourier transform*), conforme ilustrado na Fig. 2. O sinal resultante passa por um banco de filtros na escala Mel. Esta escala representa a percepção das variações em frequência pela audição humana. As frequências centrais do banco de filtros são relacionadas com as frequências em escala linear (Hz) através da expressão:

$$f_{Mel} = 1127 \cdot \ln\left(1 + \frac{f_{Hz}}{700}\right) \qquad (1)$$

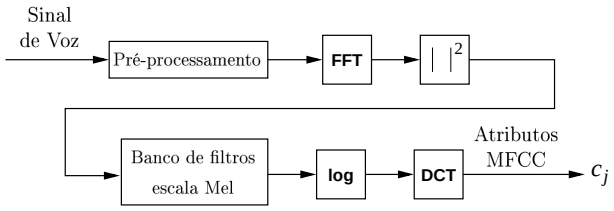

Fig. 2. A extração dos atributos MFCC.

Os coeficientes MFCC (c_j) são então obtidos pela transformada cosseno discreta (DCT - *discrete cosine transform*),

$$c_j = \sum_{k=1}^{F} (\log\, S_k)\, \cos\left[j\left(k - \frac{1}{2}\right)\frac{\pi}{F}\right]\,, \; j = 1, \ldots, D\,, \quad (2)$$

onde S_k são as potências de saída dos filtros, F é o número de filtros utilizados na escala Mel, e D é o número de coeficientes MFCC. Desta forma, de cada quadro do sinal de voz, é extraído um vetor $\vec{x} = [c_1, \ldots, c_D]^T$ de atributos de dimensão $D \times 1$. Considerando o sinal de voz composto por Q quadros, ao final da etapa de extração, a matriz de atributos é formada pelos Q vetores de atributos obtidos,

$$X = [x_1, x_2, \ldots, x_Q]\,. \qquad (3)$$

B. Os Coeficientes Dinâmicos

Os coeficientes dinâmicos ou coeficientes de diferenças, Δ, são utilizados para captar informações dinâmicas e remover características espectrais invariantes no tempo do sinal de voz [11]. Os coeficientes Δ são obtidos pelas diferenças entre vetores de atributos distanciados de uma quantidade W de quadros,

$$\Delta\vec{x_i} = \vec{x_i} - x_i\vec{}_{-W} \quad , i = 1, 2, \ldots, Q\,, \qquad (4)$$

e são, geralmente, utilizados juntamente com os vetores originais.

C. O Classificador GMM

O modelo GMM (λ) [11] é definido como uma soma ponderada de M componentes gaussianas,

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x})\,, \qquad (5)$$

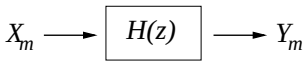

Fig. 3. Obtenção da sequência $\{Y_m\}$, de espectro colorido, a partir da filtragem de uma sequência $\{X_m\}$, de espectro branco.

onde $\vec{x}$ é um vetor de atributos com D elementos, p_i ($i = 1, 2, \cdots, M$) são os pesos das componentes, e $b_i(\vec{x})$ são componentes gaussianas com vetor média $\vec{\mu}_i$ e matriz covariância K_i. Assim, cada componente do GMM é representada por

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{\det K_i}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T K_i^{-1}(\vec{x}-\vec{\mu}_i)\right). \quad (6)$$

Desta forma, o modelo GMM do locutor é completamente representado pelos pesos, vetores média e matrizes covariância. Ou seja,

$$\lambda = \{p_i, \vec{\mu}_i, K_i\} \qquad , i = 1, ..., M\,. \quad (7)$$

Durante a fase de treinamento, os modelos de locutores são gerados a partir da matriz $X_{D\times Q}$ de atributos, utilizando o algoritmo EM (*expectation-maximization*). O objetivo é obter o modelo λ (Eq. 7), que maximize a verossimilhança entre seus parâmetros e a matriz de atributos X,

$$\log p(X|\lambda) = \frac{1}{Q}\sum_{t=1}^{Q} \log p(\vec{x}_t|\lambda)\,. \quad (8)$$

Já na fase de teste, a decisão do sistema de identificação de locutor é baseado no critério da máxima verossimilhança. Ou seja, dada uma matriz de atributos X de teste, o locutor L identificado é aquele que maximiza a soma na Eq. 8.

III. O Treinamento em Múltiplas Condições

Na técnica de treinamento em múltiplas condições proposta neste artigo, os ruídos de adição utilizados para corromper as locuções de treinamento são montados a partir de sequências de números aleatórios, ou partições, com espectros coloridos. O método de geração destas amostras é descrito a seguir.

A. Geração de Sequências Amostrais com Espectro Colorido

A cor do espectro de um ruído é definida segundo o decaimento da densidade espectral de potência (DEP), dada por

$$S(f) \propto \frac{1}{f^\beta}, \beta \in [0, 2]\,. \quad (9)$$

Desta forma, o ruído branco é aquele cujo potência é constante ($\beta \approx 0$) ao longo do espectro de frequências. Já os espectros de cores rosa e marrom são aqueles cujos decaimentos da função DEP são de 3 dB por oitava ($\beta \approx 1$) e 6 dB por oitava ($\beta \approx 2$), respectivamente.

A Fig. 3 ilustra a obtenção de uma sequência de espectro colorido $\{Y_m\}$ a partir da filtragem de um sinal $\{X_m\}$, de espectro branco.

Para que a resposta em frequência do filtro utilizado seja proporcional a $1/f^\beta$, a função de transferência $H(z)$ é escolhida como a forma proposta por Al-Alaoui [8], elevada ao expoente de ordem fracionária $\beta/2$,

$$H(z) = \left[\frac{7T}{8}\frac{(1+z^{-1}/7)}{(1-z^{-1})}\right]^{\beta/2}, \quad (10)$$

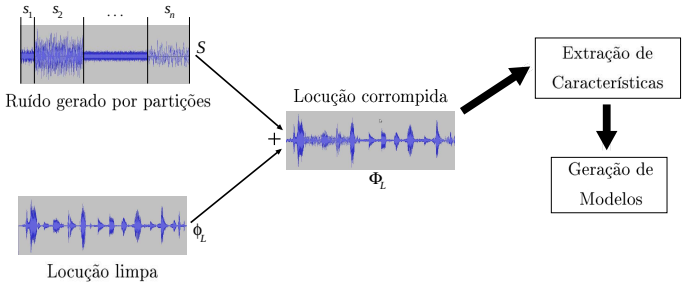

Fig. 4. Obtenção do modelo do locutor L a partir de ruídos montados em múltiplas partições.

onde T é o período de amostragem e $0 \leq \beta/2 \leq 1$.

A função DEP da sequência $\{Y_m\}$ possui a forma [12]

$$S_Y(f) \propto \left[\frac{\frac{50}{49} + \frac{2}{7}\cos(\pi f T)}{2\text{sen}(\pi f T)}\right]^\beta. \quad (11)$$

Pela Eq. 11, é possível verificar que $S_Y(f) \propto 1/f^\beta$ à medida que $f \to 0$. Para o cálculo dos coeficientes do filtro, é utilizada a expansão em série de potências da função $H(z)$ (Eq. 10). A expansão é truncada em um número finito de coeficientes, resultando em um filtro de resposta finita ao impulso (FIR - *finite impulse response*).

B. Montagem dos Ruídos de Adição

Seja ϕ_L um sinal de voz limpo captado para treinamento do modelo de um locutor L, com duração de T segundos e amostrado a uma taxa de r amostras por segundo. Para a montagem do ruído de adição, primeiramente são geradas sequências de amostras $s_i, (i = 1, 2, \cdots, n)$. Cada uma destas partições s_i é obtida com um total de $\pi_i rT$ amostras e valores distintos do expoente da DEP no intervalo $\beta \in [0, 2]$. Os coeficientes π_i, representam a ponderação, ou tamanho, de cada uma das partições na obtenção do ruído de adição, com $\sum_{i=1}^{n} \pi_i = 1$. O ruído de adição resultante é obtido pela concatenação das sequências s_i,

$$S = [\, s_1 \,|\, s_2 \,|\, \cdots \,|\, s_n \,]\,, \quad (12)$$

e possui o mesmo número de amostras da locução de treinamento limpa ϕ_L:

$$\sum_{i=1}^{n} \pi_i rT = rT \sum_{i=1}^{n} \pi_i = rT\,. \quad (13)$$

Uma nova locução corrompida de treinamento Φ_L é obtida pela soma da locução limpa ϕ_L com o ruído S gerado artificialmente. Ou seja,

$$\Phi_L = \phi_L + S\,. \quad (14)$$

Após a extração da matriz de atributos desta locução, é obtido um modelo GMM (λ_L) com ruído gerado por partições. A Fig. 4 ilustra a montagem do ruído de adição e a consequente geração dos modelos de locutores.

IV. EXPERIMENTOS E RESULTADOS

Esta Seção apresenta os resultados obtidos nos experimentos de identificação de locutor para validar a proposta de treinamento em múltiplas condições com ruídos coloridos. Os experimentos foram também avaliados utilizando o modelo GMM convencional (com locuções limpas) e obtidos pelo treinamento em múltiplas condições com locuções corrompidas por ruído branco.

A. Base de Voz

A base de voz KING foi adotada em todos os experimentos de identificação de locutor conduzidos neste trabalho. Para estes experimentos, foram utilizadas cinco sessões (1 a 5) de conversação de 49 locutores do sexo masculino. As locuções utilizadas foram todas captadas em um mesmo local e com o mesmo microfone. De cada locução, foram excluídos os períodos de silêncio, resultando em uma duração média de 20 segundos por sessão. Para cada locutor, três sessões foram utilizadas para treinamento e duas para testes. Assim, cada treinamento foi realizado com duração de 60 segundos e ficaram disponíveis 40 s de locuções para testes. Todos os experimentos de identificação apresentados neste trabalho são independentes do texto.

B. Base de Ruídos

Os quatro ruídos utilizados para corromper os sinais de voz de teste foram coletados de fontes sonoras distintas, conforme descrições apresentadas na Tab. I. O ruído AK 47 foi extraído da base FreeSFX [13], enquanto os demais são oriundos da base NOISEX-92 [14]. Antes de serem adicionados aos sinais de voz, os ruídos foram re-amostrados à taxa de 8 kHz e tiveram suas durações reduzidas a 40 s, obtendo assim a mesma configuração das locuções de teste da base de voz.

C. Ambiente de Testes

Para os experimentos de identificação de locutor, as locuções de voz da base KING disponíveis para testes foram corrompidas com os ruídos acústicos com valores de RSR de 10 dB, 15 dB e 20 dB. Este cenário foi adotado para avaliar o sistema em diversas condições de descasamento entre treinamento e teste.

Os experimentos foram realizados com locuções de teste de duração de 5 s e 1 s. Assim, os experimentos de duração de 5 s são compostos de um total de 392 testes (49 locutores $\times$ 8 locuções/locutor), enquanto os de 1 s totalizam 1960 testes (49 locutores $\times$ 40 locuções/locutor). Os erros de precisão dos resultados de identificação são de 0,2550 e 0,0510 para testes de 5 s e 1 s, respectivamente. Estes valores foram obtidos pela desigualdade de Chebyshev para um grau de confiança de 95%.

Cada uma das locuções foi dividida em quadros de 20 ms. Quadros consecutivos foram obtidos a cada 10 ms (ou seja, com 50% de sobreposição). Na etapa de extração de atributos, 20 MFCC [11] são obtidos de cada quadro, com os filtros na escala Mel em toda a faixa de frequências 0 - 4 kHz. Os experimentos também utilizaram os coeficientes dinâmicos Δ. Logo, para cada locução de treinamento e teste, foi obtida uma matriz de atributos formada por vetores coluna de 40 coeficientes (20 MFCC + 20 Δ).

TABELA I

OS QUATRO RUÍDOS ACÚSTICOS UTILIZADOS NOS EXPERIMENTOS.

Ruído	Descrição
AK 47	Sequência de disparos de um fuzil *AK 47*
Balbúrdia	100 pessoas conversando numa sala
Fábrica	Ruído de uma fábrica de equipamentos elétricos
Maq Navio	Sala de máquinas de um navio *Destroyer*

TABELA II

TAXAS DE ACERTOS (%) COM A TÉCNICA DE TREINAMENTO EM MÚLTIPLAS CONDIÇÕES PROPOSTA PARA TESTES DE 5 S.

SNR do Treinamento	SNR dos Testes			Média
	10 dB	15 dB	20 dB	
10 dB	45,58	58,67	68,28	57,51
15 dB	54,27	65,75	73,72	64,58
20 dB	57,33	71,62	79,53	69,49

Para avaliação do desempenho da técnica de treinamento em múltiplas condições proposta, foram realizados três conjuntos de experimentos descritos a seguir. Todos os modelos foram obtidos com 32 componentes gaussianas.

1) GMM Conv: Nestes experimentos, os modelos GMM dos locutores são obtidos da maneira convencional [11], ou seja, a partir das locuções de treinamento sem qualquer acréscimo de ruídos.
2) Mult Branco: No segundo conjunto de experimentos, os modelos de locutores são obtidos com o uso da técnica de treinamento em múltiplas condições utilizando ruído gaussiano branco. Para isto, múltiplas cópias das locuções de treinamento são corrompidas com RSR variando de 10 dB a 20 dB, em intervalos de 2 dB [5]. As locuções corrompidas de cada locutor são então concatenadas com a locução limpa original, e utilizadas para a obtenção do modelo GMM.
3) Mult Color: O terceiro conjunto de experimentos utiliza os modelos gerados de acordo com a proposta apresentada neste trabalho. Para isto, foram geradas três partições de 20 s de duração, resultando num sinal ruidoso com a mesma duração das locuções de treinamento. As partições foram geradas com $\beta = 0$, $\beta = 1$ e $\beta = 2$, representando assim os espectros de cores branca, rosa e marrom. Os ruídos foram adicionados às locuções de treinamento utilizando um único valor de RSR.

D. Resultados Obtidos

Para a definição da relação sinal-ruído a ser utilizada na técnica proposta (Mult Color), os ruídos coloridos foram inicialmente adicionados às locuções de treinamento com RSR de 10 dB, 15 dB e 20 dB. A Tab. II apresenta as taxas médias de acertos obtidas na identificação de locutor para testes de 5 s de duração. Como pode-se observar, o melhor desempenho foi obtido para RSR de 20 dB. Assim, este valor foi adotado nos demais experimentos Mult Color apresentados neste trabalho.

A Tab. III apresenta as taxas de acertos obtidas nos três conjuntos de experimentos de identificação de locutor para testes de 5 s e submetidos aos diferentes ruídos. Os resultados mostram que, de um total de 12 situações de ruídos consideradas nos testes (4 ruídos $\times$ 3 valores de RSR), a técnica Mult

TABELA III

TAXAS DE ACERTOS (%) PARA TESTES DE DURAÇÃO DE 5 S.

Ruído	RSR (dB)	GMM Conv	Mult Branco	Mult Color
Sem ruído		91,58	88,01	85,20
AK 47	20	80,87	80,10	84,95
	15	74,23	79,85	80,61
	10	56,89	73,21	74,23
	Média	70,66	77,72	79.93
Balbúrdia	20	83,93	75,77	80,87
	15	74,23	69,64	75,26
	10	54,08	56,38	59,44
	Média	70,75	67,26	71,85
Fábrica	20	83,16	75,77	83,42
	15	70,66	70,41	80,36
	10	46,17	58,16	63,01
	Média	66,67	68,11	75,60
Maq Navio	20	61,48	60,71	68,88
	15	44,13	48,47	50,26
	10	25,26	33,93	32,65
	Média	43,62	47,70	50,60

Color proposta neste trabalho apresenta o melhor desempenho em dez experimentos. O aumento nas taxas de acertos, em comparação com a técnica Mult Branco, chega a 9,95% para o ruído Fábrica e RSR de 15 dB. Em relação ao GMM Conv, o acréscimo na acurácia da identificação atinge 17,34% para o ruído AK 47 e RSR de 10 dB.

A Fig. 5 apresenta as médias das taxas de acertos resultantes dos experimentos para cada tipo de ruído acústico. Estes resultados são referentes a testes com duração de 5 s e 1 s. Note que a técnica Mult Color apresenta o melhor desempenho médio para os quatro ruídos considerados em testes de 5 s. Mesmo para o ruído Balbúrdia, situação em que a técnica Mult Branco não consegue melhorar o desempenho em relação ao GMM Conv, a técnica proposta obtém as maiores taxas de acertos. Para testes de 1 s, a técnica proposta apresenta os melhores resultados para três dos quatro ruídos. Conforme mostra a Fig. 5, o ruído Balbúrdia é o único caso que o modelo GMM convencional consegue desempenho semelhante ao Mult Color. Ou seja, os resultados demonstram que a técnica proposta melhora o desempenho da identificação de locutor em diferentes condições ruidosas e para testes com durações distintas. Adicionalmente, é importante ressaltar que a técnica Mult Branco utiliza as locuções de treinamento corrompidas com seis níveis distintos de ruídos, além das locuções limpas. Por outro lado, um único valor de RSR é utilizado na técnica Mult Color proposta.

V. CONCLUSÕES

Neste trabalho, é proposta uma técnica de treinamento em múltiplas condições para reconhecimento automático de locutor. Com o objetivo de diminuir o descasamento entre as fases de treinamento e teste, as locuções disponíveis para treinamento são corrompidas por um ruído gerado artificialmente. Para a montagem deste ruído, é proposta a geração de sequências de amostras com espectros coloridos distintos. Cada uma destas partições é obtida pela filtragem de uma sequência de espectro branco. A função de transferência de Al-Alaoui é utilizada para obter a função densidade espectral de potência desejada.

O desempenho da técnica de treinamento em múltiplas condições com ruídos coloridos foi avaliado para a

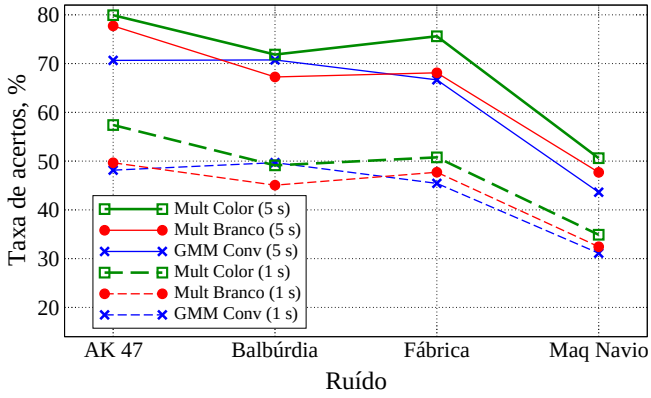

Fig. 5. Média das taxas de acertos para os diferentes tipos de ruídos acústicos.

identificação de locutor. Para o treinamento, sequências de espectro branco, rosa e marrom foram utilizadas para compor o ruído de adição. Para referência das taxas de acertos de identificação, os experimentos também foram realizados considerando modelos obtidos a partir de locuções limpas e corrompidas por ruído branco. As locuções de teste, com duração de 5 s e 1 s, foram corrompidas por quatro ruídos acústicos ambientais, extraídos de diferentes fontes, com três diferentes valores de RSR. Os resultados demonstraram que a técnica proposta neste trabalho apresenta o melhor desempenho para dez das doze situações de ruídos consideradas nos testes de 5 s de duração. Além disso, a técnica proposta apresentou as maiores taxas médias de acertos para os quatro ruídos utilizados nas locuções de testes.

REFERÊNCIAS

[1] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[2] J. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437–1461, September 1997.

[3] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," vol. 12, pp. 705–708, April 1987.

[4] D. David Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the 6th International Conference on Spoken Language Processing*, pp. 29–32, 2000.

[5] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1711–1723, July 2007.

[6] M. Keshner, "$1/f$ noise," *Proceedings of the IEEE*, vol. 70, pp. 212–218, March 1982.

[7] R. Voss and J. Clarke, "$1/f$ noise in music: Music from $1/f$ noise," *J. of the Acoustical Society of America*, vol. 63, no. 1, pp. 258–263, 1978.

[8] M. Al-Alaoui, "Novel digital integrator and differentiator," *Electronics Letters*, vol. 29, pp. 376–378, February 1993.

[9] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 259–272, April 1981.

[10] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, pp. 93–96, April 1983.

[11] D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–82, 1995.

[12] Y. Ferdi, A. Taleb-Ahmed, and M. Lakehal, "Efficient generation of $1/f^{\beta}$ noise using signal modeling techniques," *IEEE Transactions on Circuits and Systems*, vol. 55, pp. 1704–1710, July 2008.

[13] FreeSFX, "www.freesfx.co.uk."

[14] A. Varga and H. Steeneken, "Assessment for automatic speech recognition ii: Noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communications*, vol. 12, no. 3, pp. 247–251, 1993.