Realce EMDF e Treinamento em Múltiplas Condições Acústicas para Identificação de Locutor Robusta a Ruídos Não-Estacionários

L. Zão e R. Coelho

***Resumo*— Este trabalho investiga o uso do treinamento em múltiplas condições (TMC) acústicas para aprimorar a tarefa de identificação de locutor em ambientes de ruídos acústicos não-estacionários. A principal contribuição é a adoção do TMC juntamente com um método de realce de voz baseada na decomposição em modos empíricos. As técnicas de TMC adotadas neste trabalho são baseadas no uso de ruídos de espectros branco e coloridos. Experimentos de identificação automática de locutor são realizados com locuções de teste corrompidas por ruídos acústicos provenientes de diferentes fontes, em diversas relações sinal-ruído. Os resultados demonstram que o TMC aumenta a robustez da identificação de locutor em comparação ao uso apenas do realce de voz.**

***Palavras-Chave*— treinamento em múltiplas condições, realce de voz, identificação de locutor, ruídos não-estacionários.**

***Abstract*— This paper investigates the multicondition training (MT) technique for improving the speaker identification in non-stationary acoustic noises. The main contribution is the adoption of MT together with a speech enhancement approach based on the empirical mode decomposition. The MT techniques adopted in this work are based on artificial noises with white and colored spectra. Speaker identification experiments are conducted with test utterances corrupted with several acoustic noises and different signal-to-noise ratios. The results show that the MT improves the robustness of the speaker identification task.**

***Keywords*— multicondition training, speech enhancement, speaker identification, non-stationary noises.**

I. INTRODUÇÃO

A autenticação de indivíduos pela voz é considerada uma interessante solução para aplicações na área de segurança e defesa [1] [2]. A voz é considerada uma das características biométricas mais naturais para reconhecer uma pessoa. Além disso, o sinal de voz é de fácil aquisição e seu processamento pode ser considerado simples para a tecnologia atual. Sistemas de reconhecimento automático de locutor (RAL) têm ampla aceitação em aplicações como o controle de acesso, segurança da informação e investigações forenses [3].

Na literatura, os sistemas de RAL baseados nos coeficientes mel-cepestrais (MFCC - *mel-frequency cepstral coefficients*) [4] e modelos de misturas Gaussianas (GMM - *Gaussian mixture models*) [5] atingem altas taxas de acertos quando os sinais de voz são captados em ambientes limpos, i.e., sem ruídos [6]. Contudo, a ocorrência de ruídos acústicos pode levar a drásticas quedas de desempenho nestes sistemas [7]. Este impacto é atribuído à variabilidade ou ao desconhecimento das características das diferentes fontes de ruídos acústicos.

Uma técnica interessante para prover robustez ao RAL em ambientes ruidosos é o treinamento em múltiplas condições (TMC) [8] [7] [9]. 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 [7], o TMC foi implementado utilizando o ruído Gaussiano branco considerando diversos valores para a relação sinal-ruído (RSR). Os autores argumentam que o ruído branco foi escolhido devido ao desconhecimento das características dos ruídos presentes nos sinais de voz. Em [9], ruídos artificiais de espectro colorido foram utilizados para corromper as locuções de treinamento com um único valor de RSR. O treinamento em múltiplas condições com ruídos coloridos (TMCC) [9] apresentou os melhores resultados quando comparados aos obtidos com TMC com ruído branco (TMCB).

Em [10], duas técnicas de realce de sinais de voz foram avaliadas como possíveis alternativas para tornar robusta a tarefa de identificação de locutor em situações de ruídos. Na primeira abordagem (IMCRA - *improved minima controlled recursive averaging*) [11], utilizou-se a transformada de Fourier para estimar as componentes espectrais do ruído e então eliminá-las na reconstrução do sinal de voz. Na segunda, aplicou-se a decomposição em modos empíricos (EMD - *empirical mode decomposition*) [12] ao sinal de voz ruidoso para eliminar as componentes do ruído no domínio do tempo [13]. Os resultados em [10] demonstraram que o realce baseado no método EMD (EMDF - *EMD-based filtering*) aumentou a acurácia da identificação de locutor para ruídos acústicos não-estacionários.

Neste trabalho, a técnica TMCC é avaliada em conjunto com o realce EMDF para a tarefa de identificacão de locutor. O objetivo é prover maior robustez para situações de ruídos acústicos não-estacionários quando comparado à utilização apenas do realce EMDF [10]. Os experimentos de identificação de locutor são realizados considerando locuções corrompidas por seis ruídos acústicos reais, adicionados com valores de RSR entre -5 dB e 20 dB. Para o TMCC, os ruídos de espectro colorido são adicionados às locuções de treinamento considerando valores de RSR de 15 dB e 20 dB. Como base de comparação, a técnica TMCB também é adotada nos experimentos. Os resultados demonstram que o TMCC aumenta a taxa média de acertos quando comparado ao uso do realce EMDF. A acurácia média obtida com o TMCC também

L. Zão e R. Coelho, Laboratório de Processamento de Sinais Acústicos, Departamento de Engenharia Elétrica, Instituto Militar de Engenharia (IME), Rio de Janeiro, Brasil, E-mails: zao@ime.eb.br, coelho@ime.eb.br. Este trabalho foi parcialmente financiado pelo CNPq 304254/2012-6.

é maior em relação àquela obtida com o TMCB.

O restante deste trabalho está organizado da seguinte forma. A Seção II apresenta a técnica de realce de voz EMDF. Na Seção III, são descritos os principais conceitos referentes à identificação automática de locutor, incluindo o treinamento em múltiplas condições com ruídos coloridos. A descrição dos experimentos, da base de ruídos utilizada assim como os resultados obtidos com as técnicas TMC e EMDF são apresentados na Seção IV. Finalmente, a Seção V conclui o presente trabalho.

II. Realce da Voz

A técnica EMDF foi proposta em [13] para realçar sinais de voz corrompidos por ruídos de baixas frequências. Nesta técnica, o método EMD é inicialmente aplicado sobre o sinal ruidoso resultando em um conjunto de funções intrínsecas de modo (IMF - *intrinsic mode functions*). Em seguida, utiliza-se um algoritmo para determinar quais das IMFs resultantes são predominantemente compostas por ruído. Finalmente, estas são removidas para a recomposição do sinal de voz limpo.

A. EMD

O método EMD [12] é uma forma não-linear de análise de sinais não-estacionários. Considere um sinal $y(t)$ contendo dois máximos locais consecutivos nos pontos t_- e t_+. Para valores de t no intervalo $t_- \leq t \leq t_+$, pode-se definir uma componente de altas frequências do sinal que passa por estes máximos e pelo mínimo local que existe entre eles. Desta componente, chamada de detalhes $d(t)$, identifica-se uma componente de tendência local $m(t)$, tal que $y(t) = d(t) + m(t)$ no intervalo $t_- \leq t \leq t_+$. Uma IMF é definida pelo conjunto das componentes de detalhes, quando esta decomposição é aplicada a todas as oscilações presentes no sinal $y(t)$. Analogamente, um sinal residual é definido pelo conjunto de componentes de tendência locais. Aplicando repetidamente o procedimento sobre o sinal residual, chega-se a um conjunto de IMFs e a um resíduo de baixas frequências.

O algoritmo para o método EMD aplicado sobre um sinal $y(t)$ pode ser dividido nos seguintes passos [12] [14]:

1) Identificar todos os extremos (máximos e mínimos locais) de $y(t)$;
2) Obter as envoltórias $e_{max}(t)$ e $e_{min}(t)$, utilizando interpolação por *splines* cúbicas nos pontos de máximo e mínimo, respectivamente;
3) Calcular a componente de tendências como a média entre as envoltórias: $m(t) = (e_{min}(t) + e_{max}(t))/2$;
4) Extrair os detalhes: $d(t) = y(t) - m(t)$;
5) Repetir a iteração sobre o sinal residual $m(t)$.

Em geral, para garantir que a componente de detalhes $d(t)$ extraída no passo (4) seja considerada uma IMF, os passos (1)-(4) são repetidos com $d(t)$ no lugar de $y(t)$. Este processo é repetido até garantir que a nova função $d(t)$ tenha média próxima de zero. Ao final de um número finito N de iterações, o sinal $y(t)$ pode ser escrito como

$$y(t) = \sum_{n=1}^{N} \text{IMF}_n(t) + m(t)\,, \qquad (1)$$

onde $\text{IMF}_n(t), 1 \leq n \leq N$, são as funções de detalhes obtidas no passo (4) de cada iteração, e $m(t)$ é o sinal residual obtido na última iteração.

B. Técnica EMDF

Na proposta da técnica EMDF [13], para determinar quais IMFs devem ser excluídas na reconstrução do sinal de voz, utiliza-se o fato de que a maior parte da energia de um sinal de voz limpo se concentra nas quatro primeiras IMFs. Os autores demonstram que as variâncias dos modos decaem significativamente para índices $n \geq 5$. Assim, o aumento na variância destes modos, isto é, $Var[\text{IMF}_n(t)] > Var[\text{IMF}_{n-1}(t)], n > 4$, indica que estes são fortemente afetados pelas componentes de baixas frequências dos ruídos.

Para a filtragem, o sinal de voz é primeiramente dividido em pequenos quadros e, em cada um destes, é efetuada a busca pelas IMFs mais comprometidas por ruídos. Esta busca quadro a quadro é necessária já que, para sinais não-estacionários, as características do ruído podem se alterar ao longo do tempo. O algoritmo do realce EMDF pode então ser resumido nos seguintes passos [13]:

1) Dividir o sinal de voz $y(t)$ em quadros $y_l(t), l = 1, \ldots, Q$;
2) Para cada quadro $y_l(t)$, efetuar a decomposição em N funções $\text{IMF}_n(t), n = 1, \ldots, N$;
3) Estimar as variâncias $V_l(n) = Var\,[\text{IMF}_n(t)]$;
4) Identificar os índices dos picos p_l tais que $V(p_l) > V(p_l - 1)$ e $V(p_l) > V(p_l + 1)$, para $p_l > 4$;
5) Determinar o índice v_l do vale imediatamente anterior ao pico p_l, isto é, $V(v_l) < V(v_l - 1)$ e $V(v_l) < V(v_l + 1)$, para $v_l < p_l$;
6) Determinar para qual quadro $\hat{l}$ a diferença $p_{\hat{l}} - v_{\hat{l}}$ é máxima;
7) Reconstruir o sinal de voz $\hat{y}(t) = \sum_{n=1}^{N'} \text{IMF}_n(t)$, onde $N' = v_{\hat{l}}$.

III. Identificação Automática de Locutor

Um sistema de identificação automática de locutor é geralmente dividido em duas fases: treinamento e testes. 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 e o sistema decide a qual dos usuários cadastrados pertence a locução de teste.

Na literatura, os coeficientes mel-cepestrais [4] e o modelo de misturas gaussianas [5] são considerados referência de bom desempenho em sistemas de identificação de locutor.

A. Extração dos 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*). 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:

$$f_{Mel} = 1127 \cdot \ln\left(1 + \frac{f_{Hz}}{700}\right) \tag{2}$$

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

$$c_h = \sum_{k=1}^{F} (\log\ S_k)\ \cos\left[h\left(k - \frac{1}{2}\right)\frac{\pi}{F}\right]\ ,\ h = 1, \ldots, D\,, \tag{3}$$

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,

$$X = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_Q]\ . \tag{4}$$

B. Classificador GMM

O modelo GMM (λ) [5] é definido como uma soma ponderada de M componentes gaussianas,

$$p(\vec{x}|\lambda) = \sum_{j=1}^{M} p_j b_j(\vec{x}) \tag{5}$$

onde $\vec{x}$ é um vetor de atributos com D elementos, p_j ($j = 1, 2, \cdots, M$) são os pesos das componentes, e $b_j(\vec{x})$ são componentes gaussianas com vetor média $\vec{\mu}_j$ e matriz covariância K_j. Desta forma, cada componente do GMM é representada por

$$b_j(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{\det\ K_j}}\ \exp\!\left(-\frac{1}{2}(\vec{x} - \vec{\mu}_j)^T K_j^{-1}(\vec{x} - \vec{\mu}_j)\right). \tag{6}$$

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

$$\lambda = \{p_j, \vec{\mu}_j, K_j\} \qquad , j = 1, ..., M\,. \tag{7}$$

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 em (7), que maximize a verossimilhança entre seus parâmetros e a matriz de atributos X,

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

Já na fase de teste, a decisão do sistema de identificação de locutor é baseada no critério da máxima verossimilhança. Ou seja, dada uma matriz de atributos X de teste, o locutor $\hat{L}$ identificado é aquele cujo modelo maximiza a soma em (8),

$$\hat{L} = \arg\max_k \log P(X|\lambda_k) = \arg\max_k \sum_{t=1}^{Q} \log p(\vec{x}_t|\lambda)\,. \tag{9}$$

C. Treinamento em Múltiplas Condições

Seja Φ_L^0 a locução limpa disponível para treinamento do locutor L. Um conjunto de locuções em múltiplas condições acústicas (Φ_L^i, $i = 1, 2, \ldots, R$) é obtido a partir da adição de R ruídos artificiais de espectros coloridos a Φ_L^0. As matrizes de atributos, extraídas de cada uma das locuções Φ_L^i, são então utilizadas para obter um conjunto de R modelos (λ_S^i) para o locutor L:

$$p(\vec{x}|\lambda_L^i) = \sum_{j=1}^{M} p_j^i\ b_j^i(\vec{x}), \qquad i = 1, \ldots, R\,. \tag{10}$$

De acordo com (10), cada GMM λ_L^i é composto por M densidades Gaussianas. Assim, um total de $R \times M$ componentes são geradas e armazenadas para cada locutor L. Em analogia a (7), os R modelos referentes a L são parametrizados por

$$\lambda_L^i = \{p_j^i, \vec{\mu}_j^i, K_j^i \,|\, j = 1, \ldots, M\}\,, \qquad i = 1, \ldots, R\,. \tag{11}$$

O modelo treinado em múltiplas condições (Λ_L) é então definido [9] pela coleção de todos os parâmetros estimados em (11),

$$\Lambda_L = \bigcup_{i=1}^{R} \lambda_L^i = \{p_j^i, \vec{\mu}_j^i, K_j^i \,|\, i = 1, \ldots, R\,;\, j = 1, \ldots, M\,\}\,. \tag{12}$$

Considerando os modelos Λ_L, a regra de decisão adotada na tarefa de identificação de locutor é adaptada para

$$\hat{L} = \arg\max_L \sum_{t=1}^{Q} \log p(\vec{x}_t|\Lambda_L)\,, \tag{13}$$

onde $p(\vec{x}|\Lambda_L)$ é ajustada para considerar todas as componentes Gaussianas armazenadas em Λ_L, ou seja,

$$p(\vec{x}|\Lambda_L) = \sum_{i=1}^{R}\sum_{j=1}^{M} \pi_i p_j^i\ b_j^i(\vec{x})\,. \tag{14}$$

Cada termo π_i em (14) representa o peso de uma condição de ruído Φ_L^i, com $\sum_{i=1}^{R} \pi_i = 1$. Assim como em [9], assumindo que não há qualquer conhecimento prévio sobre as estatísticas dos ruídos presentes nas locuções de testes, os experimentos realizados neste trabalho utilizam os valores $\pi_i = 1/R$ para cada $i = 1, \ldots, R$.

IV. Experimentos de Identificação de Locutor

Os experimentos de identificação de locutor foram realizados utilizando um subconjunto de 168 locutores da base de voz TIMIT [15]. Para análise de desempenho, foram utilizadas 336 locuções de teste, sendo duas por locutor, com duração média de 3 segundos e taxa de amostragem de 16 kHz. Para o treinamento dos modelos de cada locutor, foram utilizadas oito locuções limpas, diferentes daquelas utilizadas nos testes.

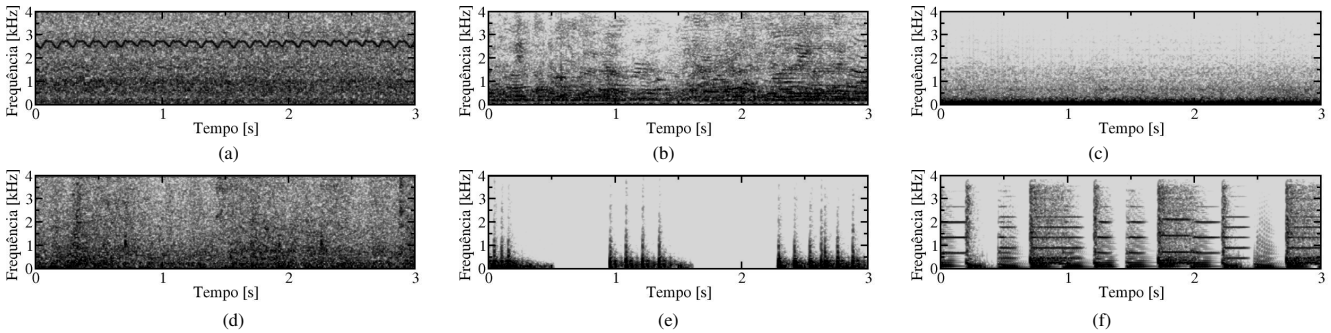

Fig. 1. Espectrogramas dos ruídos acústicos utilizados neste trabalho: (a) Avião, (b) Balbúrdia, (c) Carro, (d) Fábrica, (e) Metralhadora e (f) *Ringtone* [10].

A. Base de Ruídos

Para os testes, cada uma das 336 locuções foi corrompida por seis ruídos acústicos reais: Avião, Balbúrdia, Carro, Fábrica, Metralhadora e *Ringtone*. Com exceção do ruído *Ringtone*, obtido em [16], todos os ruídos foram extraídos da base NOISEX-92 [17]. Antes de serem adicionados, os ruídos foram subamostrados para a mesma taxa dos sinais de voz.

A Fig. 1 ilustra os espectrogramas de segmentos de 3 segundos de duração para cada um dos ruídos. Como pode-se verificar, alguns dos ruídos possuem acentuadas variações temporais nos seus espectros (ruído Matralhadora, por exemplo). Em [10], foi demonstrado que os ruídos Balbúrdia, Fábrica, Metralhadora e *Ringtone* são não-estacionários, com diferentes índices de não-estacionariedade (INS - *index of nonstationarity*) [18]. Já os ruídos Avião e Carro são predominantemente estacionários.

B. Resultados com o Realce EMDF

Para os experimentos de identificação de locutor, as locuções foram divididas em quadros de 20 ms, com 50% de sobreposição. De cada quadro, foi extraído um vetor de coeficientes MFCC com 12 componentes. Para a modelagem dos locutores, foi adotado o classificador GMM com 32 componentes Gaussianas.

A Tab. I apresenta as taxas de acertos da identificação de locutor com as locuções corrompidas da base TIMIT. Na Tab. II, os resultados correspondem aos sinais de voz realçados pelo método EMDF antes de serem aplicados à identificação de locutor. Como pode-se observar, a filtragem EMDF aplicada sobre os sinais de voz ruidosos conseguiu aumentar a taxa média de acertos. O melhor resultado foi obtido para o ruído Carro, com um aumento nas taxas de acertos de 55,61% para 83,23%. Note ainda que o realce obteve as maiores taxas para as condições de ruído mais severas, inclusive para os ruídos com maiores variações nos espectros de frequências (Metralhadora e *Ringtone*). Na média, a técnica EMDF aumentou a taxa de acertos de 46,92% para 50,30%, representando um incremento médio de 3,38%.

C. Resultados com Realce e TMC

Para a avaliação da técnica TMCC, apresentada na Seção III-C, são utilizados três ruídos artificiais obtidos segundo o gerador proposto em [19], com padrão Gaussiano e densidade espectral de potência (DEP) definida por $\mathcal{S}(f) \propto 1/f^{\beta}$. Para representar as variadas formas da DEP de ruídos acústicos reais, foram considerados espectros de cores branca ($\beta = 0$), rosa ($\beta = 1$) e marrom ($\beta = 2$) para os ruídos gerados. Assim, os modelos obtidos com a técnica TMCC são compostos por $3 \times 32 = 96$ componentes Gaussianas. Dois conjuntos de experimentos são realizados considerando valores distintos de RSR: 15 dB e 20 dB.

A técnica de TMC com ruído branco (TMCB) é também implementada, como base de comparação para o TMCC. Seguindo o procedimento definido em [7], múltiplas cópias das locuções de treinamento são inicialmente obtidas corrompendo-se o sinal limpo com o ruído Gaussiano branco nos seguintes valores de RSR: 10, 12, 14, 16, 18 e 20 dB. Finalmente, as locuções limpas e corrompidas são concatenadas e, a partir destas, um modelo GMM com 128 componentes é obtido para cada locutor. Nas duas técnicas de TMC, o realce EMDF é também aplicado às locuções de treinamento antes da extração das matrizes de atributos.

As Tabs. III e IV apresentam os resultados de identificação de locutor obtidos com a técnica TMCC para RSR de 15 dB e

TABELA I

TAXAS DE ACERTOS (%) NA IDENTIFICAÇÃO DE LOCUTOR OBTIDOS COM SINAIS DE VOZ CORROMPIDOS POR DIVERSOS RUÍDOS [10].

Ruído	RSR (dB)						
	-5	0	5	10	15	20	Média
Metralhadora	60,42	76,19	84,23	91,96	96,43	98,21	84,57
Carro	19,05	31,55	49,11	64,29	79,46	90,18	55,61
Balbúrdia	4,76	13,10	35,42	71,13	91,07	96,73	52,03
Ringtone	7,14	17,26	36,61	60,12	80,95	94,64	49,45
Fábrica	0,89	0,60	2,98	15,18	43,75	80,65	24,01
Avião	0,30	0,60	0,60	7,14	26,19	60,42	15,87
Média	15,43	23,21	34,82	51,64	69,64	86,81	46,92

TABELA II

TAXAS DE ACERTOS (%) NA IDENTIFICAÇÃO DE LOCUTOR OBTIDAS COM REALCE EMDF [10].

Ruído	RSR (dB)						
	-5	0	5	10	15	20	Média
Carro	63,99	77,98	84,23	87,50	91,37	94,35	83,23
Metralhadora	65,77	77,68	82,44	86,90	90,77	93,15	82,79
Ringtone	9,82	27,38	45,24	64,58	79,17	86,61	52,13
Balbúrdia	3,57	10,71	30,95	63,39	82,74	92,86	47,37
Fábrica	1,49	0,89	3,27	14,29	43,15	73,51	22,77
Avião	0,89	1,19	1,79	4,76	21,13	51,19	13,49
Média	24,26	32,64	41,32	53,57	68,06	81,94	50,30

TABELA III

TAXAS DE ACERTOS (%) NA IDENTIFICAÇÃO DE LOCUTOR OBTIDAS COM REALCE EMDF E A TÉCNICA TMCC COM RSR DE 15 DB.

Ruído	RSR (dB)						
	-5	0	5	10	15	20	Média
Carro	63,99	82,14	84,23	88,10	89,29	91,67	83,23
Metralhadora	53,87	70,24	78,27	85,12	87,80	90,48	77,63
Ringtone	5,65	15,18	34,52	56,55	73,21	83,63	44,79
Balbúrdia	5,06	13,99	42,26	64,58	78,87	87,20	48,66
Fábrica	1,19	3,57	19,35	42,56	71,43	81,25	36,56
Avião	0,60	2,08	5,36	16,37	42,86	68,45	22,62
Média	21,73	31,20	44,00	58,88	73,91	83,78	52,25

TABELA IV

TAXAS DE ACERTOS (%) NA IDENTIFICAÇÃO DE LOCUTOR OBTIDAS COM REALCE EMDF E A TÉCNICA TMCC COM RSR DE 20 DB.

Ruído	RSR (dB)						
	-5	0	5	10	15	20	Média
Carro	67,26	87,20	91,07	91,96	93,15	92,26	87,15
Metralhadora	63,99	76,49	84,52	91,37	92,56	94,94	83,98
Ringtone	6,25	17,86	38,39	65,18	80,06	89,29	49,50
Balbúrdia	3,57	12,80	39,58	75,00	87,50	92,86	51,88
Fábrica	2,38	2,68	15,77	46,73	75,60	89,88	38,84
Avião	1,19	1,19	3,57	17,56	47,02	72,92	23,91
Média	24,11	33,04	45,49	64,63	79,32	88,69	55,88

TABELA V

TAXAS DE ACERTOS (%) NA IDENTIFICAÇÃO DE LOCUTOR OBTIDAS COM REALCE EMDF E A TÉCNICA TMCB.

Ruído	RSR (dB)						
	-5	0	5	10	15	20	Média
Carro	60,42	78,27	83,33	89,88	92,56	94,05	83,09
Metralhadora	68,15	79,46	83,33	87,80	89,58	93,15	83,58
Ringtone	8,33	22,32	43,75	65,18	79,76	86,90	51,04
Balbúrdia	2,68	9,82	29,17	62,80	80,36	89,58	45,73
Fábrica	1,19	2,08	10,71	23,51	47,32	68,75	25,60
Avião	0,89	0,60	2,38	8,04	31,55	51,79	15,87
Média	23,61	32,09	42,11	56,20	70,19	80,70	50,82

20 dB, respectivamente. As taxas de acertos correspondentes ao TMCB estão descritas na Tab. V. Como pode-se observar, a maior acurácia média é obtida com o TMCC e RSR de 20 dB. Esta técnica apresenta ganho médio absoluto de 5,58% (de 50,30% para 55,88%) em relação à utilização apenas do realce EMDF. Note que o desempenho médio do TMCC para ambos os casos de RSR é também superior ao da técnica TMCB. Com exceção do ruído *Ringtone*, o TMCC com RSR de 20 dB apresenta as maiores taxas de acertos para todas as fontes de ruídos, tanto em relação ao TMCB quanto ao realce EMDF apenas. Mesmo para o *Ringtone*, o TMCC atinge melhor acurácia para os maiores valores de RSR. É importante ressaltar que o melhor desempenho do TMCC em relação ao TMCB é obtido mesmo com um menor número de componentes Gaussianas: 96 para cada modelo no TMCC e 128 para o TMCB.

V. CONCLUSÃO

Este trabalho investigou a utilização do treinamento em múltiplas condições acústicas em conjunto com a técnica EMDF para prover robustez a um sistema de identificação automática de locutor em situações de ruídos acústicos não-estacionários. Seis ruídos acústicos, coletados de diferentes fontes reais, foram utilizados para corromper as locuções utilizadas nos testes. Os resultados demonstraram que o TMC melhorou o desempenho da identificação de locutor. O aumento absoluto na taxa média de acertos chegou a 5,58% em relação à utilização apenas do realce EMDF. Além disso, o TMCC atingiu os melhores resultados para cinco das seis fontes de ruídos consideradas neste trabalho, incluindo ruídos estacionários e não-estacionários. Em conjunto, a técnica TMCC e o realce EMDF aumentaram a acurácia média do sistema de identificação de locutor em 8,96%, de 46,92% para 55,88%.

REFERÊNCIAS

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

[2] J. Naik, "Speaker verification: A tutorial," *IEEE Communications Magazine*, vol. 28, pp. 42–48, January 1990.

[3] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic Speaker Recognition," *IEEE Signal Processing Magazine*, vol. 26, pp. 95–103, 2009.

[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, August 1980.

[5] 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.

[6] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[7] 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.

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

[9] L. Zão and R. Coelho, "Colored noise based multicondition training technique for robust speaker identification," *IEEE Signal Processing Letters*, vol. 18, pp. 675–678, November 2011.

[10] L. Zão and R. Coelho, "Realce de sinais de voz em presença de ruídos acústicos não-estacionários utilizando o método EMD," in *Anais do XXX Simpósio Brasileiro de Telecomunicações (SBrT'12)*, pp. 1–5, Setembro 2012.

[11] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 466–475, September 2003.

[12] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, March 1998.

[13] N. Chatlani and J. Soraghan, "EMD-based filtering (EMDF) of low-frequency noise for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1158–1166, May 2012.

[14] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, pp. 112–114, February 2004.

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.

[16] FindMIDIs.com, "http://www.findmidis.com."

[17] 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.

[18] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Transactions on Signal Processing*, vol. 58, pp. 3459–3470, July 2010.

[19] L. Zão and R. Coelho, "Generation of coloured acoustic noise samples with non-Gaussian distributions," *IET Signal Processing*, vol. 6, pp. 684–688, September 2012.