Método Combinado de Multitreinamento e Realce de Sinais em Presença de Interferências Acústicas

L. Zão e R. Coelho

***Resumo*— Este trabalho avalia a combinação do realce de sinais de voz com o método de treinamento em múltiplas condições para a tarefa de identificação automática de locutor. Três propostas recentes de realce são analisadas, sendo duas delas no domínio do tempo. Os experimentos de identificação de locutor são realizados com locuções de testes corrompidas por cinco ruídos acústicos com diferentes índices de não-estacionariedade. Os resultados mostram que as técnicas de realce no domínio do tempo são capazes de aumentar a acurácia da identificação em comparação com o método de multitreinamento. A combinação de realce com multitreinamento atinge as maiores taxas de acertos mesmo para uma fonte acústica altamente não-estacionária.**

***Palavras-Chave*— identificação de locutor, realce de sinais de voz, multitreinamento, ruídos acústicos.**

***Abstract*— This paper evaluates the combination of speech enhancement techniques with multicondition training to achieve robust speaker identification in noisy environments. Three recently proposed speech enhancement techniques are compared in this work. The speaker identification experiments are conducted with speech signals corrupted by five acoustic noises with different indexes of nonstationarity. The results show that the time-domain speech enhancement approaches improve the speaker identification accuracy when compared to the adoption of the multicondition training only. The combination of these techniques achieves the highest identification scores even for the highly nonstationary acoustic noise source.**

***Keywords*— speaker identification, speech enhancement, multicondition training, acoustic noise.**

I. INTRODUÇÃO

A voz é uma das características biométricas mais naturais dos seres humanos. Além de ser de fácil aquisição, o processamento do sinal de voz é considerado relativamente simples para a tecnologia atual. Por estes motivos, sistemas de autenticação de indivíduos pela voz têm ampla aceitação para aplicações de controle de acesso, segurança de dados e investigações forenses.

Sistemas de reconhecimento automático de locutor (RAL) são implementados em duas fases: treinamento e testes. Cada uma destas fases é geralmente composta de três etapas: aquisição e pré-processamento do sinal de voz, extração de atributos da voz e classificação do locutor. Na etapa de classificação, podem ser efetuadas as tarefas de verificação ou identificação de locutor. Na identificação, o sistema decide a qual dos locutores cadastrados pertence a locução de teste. Já na verificação, o sistema decide se aceita ou não a identidade declarada pelo locutor. Na literatura, sistemas de identificação de locutor baseadas nos coeficientes mel-cepestrais (MFCC - *mel-frequency cepstral coefficients*) [1] e modelos de misturas Gaussianas (GMM - *Gaussian mixture models*) [2] alcançam bom desempenho quando os sinais de voz são captados em ambientes limpos, i.e., sem ruídos. Contudo, a ocorrência de ruídos acústicos pode levar a drásticas quedas nas taxas de acertos de identificação [3], [4].

Um método interessante para compensar a ocorrência dos ruídos acústicos em sistemas de RAL é o treinamento em múltiplas condições [3], [4], [5]. Nesta proposta, os modelos de locutor são obtidos a partir de locuções artificialmente corrompidas por ruídos acústicos, o que leva a uma redução do descasamento de condições entre as fases de treinamento e testes. Em [3], sequências de ruído Gaussiano branco foram utilizadas para corromper as locuções de treinamento com diferentes valores de razão sinal-ruído (SNR - *signal-to-noise ratio*). Já em [4], o multitreinamento (MT) foi implementado com ruídos artificiais de espectro colorido adicionados aos segmentos de voz com SNR de 20 dB. Os resultados demonstraram que o MT com ruído de espectro colorido (MTC) obteve os maiores ganhos de acurácia na identificação de locutor considerando ruídos de diferentes fontes acústicas.

Técnicas de realce têm sido propostas com o objetivo de suprimir ou reduzir os efeitos causados pelos ruídos acústicos. A maioria destas abordagens [6], [7] utiliza a transformada de Fourier de tempo curto para estimar o espectro do ruído. Estas componentes são então subtraídas ou compensadas do espectro do sinal de voz. Nos últimos anos, técnicas de realce no domínio do tempo [8], [9], [10], [11] têm sido apresentadas como alternativa aos métodos espectrais. Nas abordagens temporais, evita-se a estimação explícita das componentes espectrais dos ruídos acústicos. Isto resulta em melhores resultados de qualidade e inteligibilidade dos sinais de voz, principalmente quando corrompidos por ruídos acústicos não-estacionários.

Este artigo investiga a combinação do realce de sinais de voz com o método de multitreinamento MTC para aumentar a robustez de um sistema de RAL. No presente trabalho, esta análise é realizada considerando a tarefa de identificação de locutor. Para o realce, são abordadas três técnicas apresentadas recentemente. A proposta UMMSE (*unbiased minimum mean-square error*) utiliza o estimador definido em [6] para obter as componentes do ruído no domínio da frequência. O realce EMDH foi introduzido em [10] e utiliza a decomposição empírica de modos (EMD - *empirical mode decomposition*) [12] e o expoente de Hurst [13] para selecionar as componentes da voz mais afetadas pelo ruído. Já o realce TASE (*time-amplitude standard deviation estimator*) [11] utiliza uma adaptação do estimador DATE (*d-Dimensional Trimmed Estimator*) [14] para obter o desvio padrão do ruído acústico

L. Zão e R. Coelho*, Laboratório de Processamento de Sinais Acústicos (lasp.ime.eb.br), 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/PQ 307866/2015-7.

diretamente das amplitudes do sinal ruidoso.

A combinação do realce com o MTC é avaliado em experimentos de identificação de locutor com sinais de voz corrompidos por cinco ruídos acústicos não-estacionários. A matriz de atributos é formada pela fusão entre coeficientes MFCC e o vetor de atributos pH [15]. Os experimentos são realizados considerando valores de SNR entre 10 dB e 20 dB. As taxas de acertos são comparadas com aquelas obtidas em testes de identificação de locutor com e sem o multitreinamento. Os resultados demonstram que a combinação MT + TASE leva aos melhores resultados de identificação para três fontes acústicas de ruído. Para as demais, o melhor desempenho é alcançado pela combinação MT + EMDH.

O restante deste trabalho está organizado da seguinte forma. A Seção II descreve um sistema de identificação de locutor. Isto inclui os atributos e classificadores utilizados neste trabalho, bem como a técnica de multitreinamento. Na Seção III são brevemente descritas as três técnicas de realce: UMMSE, EMDH e TASE. Os resultados dos experimentos de identificação de locutor são apresentados e discutidos na Seção IV. Finalmente, a Seção V conclui o presente artigo.

II. Identificação de Locutor

Em um sistema de identificação de locutor, tanto a fase de treinamento quanto a fase de testes se inicia com a digitalização e o janelamento do sinal de voz em segmentos de curta duração (20-32 ms). Em seguida, vetores de características da voz são extraídos e concatenados para formar uma matriz de atributos. Na fase de treinamento, esta matriz é então utilizada para a obtenção do modelo do locutor. Já nos testes, a matriz de atributos é confrontada com os modelos previamente armazenados e o sistema decide a qual usuário pertence a locução de teste.

A. Coeficientes MFCC

O primeiro passo para a extração dos coeficientes MFCC é transformar o sinal de voz janelado 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:

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

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

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

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 de atributos $\vec{x} = [c_1, \ldots, c_D]^T$. 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,

$$X = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_Q]\ . \tag{3}$$

B. Vetor de Atributos pH

O atributo pH foi proposto em [15] e consiste em um vetor de expoentes de Hurst (H). O parâmetro de Hurst ($0 \leq H \leq 1$) expressa a dependência temporal do sinal de voz. Seja o sinal de voz representado por um processo estocástico $y(t)$, com função autocorrelação normalizada definida por

$$\rho(k) = \frac{E\left\{(y(t) - \mu_y)(y(t+k) - \mu_y)\right\}}{\sigma_y^2}\,, \tag{4}$$

onde μ_y e σ_y^2 são a média e a variância de $y(t)$, respectivamente. O expoente de Hurst ($0 \leq H \leq 1$) é definido pela taxa de decaimento da função autocorrelação normalizada, $\rho(k)$, que possui comportamento assintótico dado por

$$\rho(k) \sim H(2H-1)k^{2(H-1)}, \quad \text{quando } k \to \infty\,. \tag{5}$$

O valor de H está relacionado com as características espectrais de $y(t)$. Isto significa que $y(t)$ é predominantemente composto por altas frequências para valores $H < 1/2$. Para o caso $H = 1/2$, $S_y(f)$ é aproximadamente constante ao longo de todo o espectro de frequências, correspondendo ao ruído branco. Já para os valores de $H \in (1/2\,,\, 1]$, a maior parte da energia de $y(t)$ está concentrada nas baixas frequências.

Para os experimentos de identificação de locutor, o vetor pH é extraído com o estimador multi-dimensional baseado em *wavelets* (M-dim-wav - *multi-dimensional wavelet-based estimator*) [15]. Em [16], foi demonstrado que o vetor pH é capaz de prover robustez a um sistema de verificação de locutor sujeito a ruídos acústicos não-estacionários.

C. Classificador GMM

O modelo GMM (λ) [2] é definido como uma soma ponderada de M componentes gaussianas,

$$p(\vec{x}|\lambda) = \sum_{j=1}^{M} p_j b_j(\vec{x}) \tag{6}$$

onde $\vec{x}$ é um vetor de atributos, 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. 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 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),

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

D. Treinamento em Múltiplas Condições

No multitreinamento, a locução disponível para treinamento de cada locutor é artificialmente corrompida antes de ser utilizada para obtenção dos modelos. Assim, seja Φ_L^0 a locução limpa disponível para treinamento do locutor L. Um conjunto de locuções em múltiplas condições acústicas ($\Phi_L^i,\ i = 1, 2, \ldots, R$) 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 gerar 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 [4] pela coleção de todos os parâmetros estimados em (11), $\Lambda_L = \bigcup_{i=1}^{R} \lambda_L^i$. Considerando os modelos Λ_L, a regra de decisão adotada na tarefa de identificação de locutor é adaptada para

$$\hat{L} = \arg\max_{L} \sum_{t=1}^{Q} \log p(\vec{x}_t|\Lambda_L)\,, \tag{12}$$

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

$$p(\vec{x}|\Lambda_L) = \sum_{i=1}^{R} \sum_{j=1}^{M} \pi_i p_j^i\, b_j^i(\vec{x})\,. \tag{13}$$

Cada termo π_i em (13) representa o peso de uma condição de ruído Φ_L^i, com $\sum_{i=1}^{R} \pi_i = 1$. Assim como em [4], 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 $i = 1, \ldots, R$.

III. Técnicas de Realce de Voz

Esta seção descreve as três técnicas de realce de sinais de voz utilizadas neste trabalho: UMMSE, EMDH e TASE. Estas técnicas obtiveram interessantes resultados no realce de sinais de voz corrompidos por ruídos acústicos não-estacionários de diferentes fontes. Neste estudo, o realce é aplicado na etapa de pré-processamento, antes da segmentação do sinal de voz.

A. UMMSE

No método de realce UMMSE, o conceito de incerteza de presença de voz é introduzido ao estimador de espectro do ruído proposto em [17]. Diferentemente de outras propostas, o estimador UMMSE não necessita das estatística de vários quadros passados para estimar o espectro de potência do ruído. Isto significa que o UMMSE consegue captar as mudanças no espectro de ruídos acústicos não-estacionários com pequeno tempo de atraso. Após a estimação das componentes espectrais do ruído, o espectro do sinal de voz é obtido pela técnica baseada no filtro de Wiener [18]. A filtragem de Wiener é adotada pois, assim como o estimador UMMSE, ela assume que os coeficientes espectrais do ruído e do sinal de voz obedecem a distribuições Gaussianas.

B. EMDH

A técnica de realce EMDH[1] foi proposta em [10] para reduzir o efeito de ruídos acústicos não-estacionários no domínio do tempo. Nesta proposta, o método não-linear EMD [12] é inicialmente aplicado sobre o sinal ruidoso resultando em um conjunto de funções intrínsecas de modo (IMF - *intrinsic mode functions*). A descrição das principais características, desafios e limitações da decomposição EMD é apresentada em [19]. Após a decomposição do sinal ruidoso, a técnica EMDH divide as IMFs em segmentos de curta duração. O expoente de Hurst (H) [13] é então utilizado para identificar as componentes mais corrompidas pelo ruído acústico. Em cada quadro, são selecionadas as IMFs cujos valores de H são maiores que o limiar definido por $H_{\text{th}} = 0.9$. Os demais modos são então somados para reconstruir cada um dos quadros do sinal de voz. Finalmente, os quadros são concatenados para formar a versão completa do sinal de voz realçado.

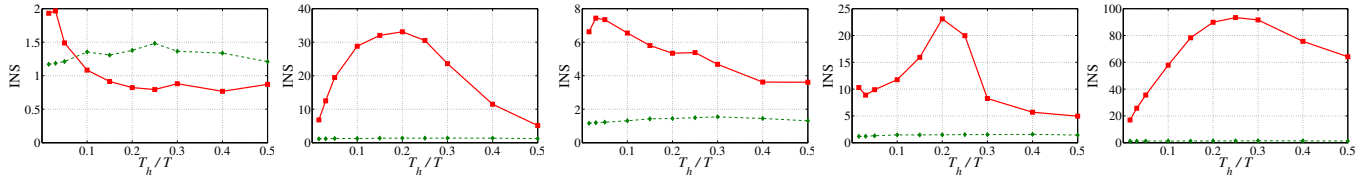
C. TASE

No método de realce TASE[1] [11], as componentes do ruído são detectadas e compensadas diretamente das amostras do sinal ruidoso. Para isto, a estimação do desvio padrão do ruído é efetuada sem utilizar qualquer forma de decomposição temporal ou análise espectral. O valor estimado do desvio padrão é utilizado para definir um limiar para a seleção das componentes para compor o sinal realçado. O realce TASE pode ser dividido em três etapas:

1) Segmentar o sinal de voz $y(t)$ em Q quadros de curta duração, denotados por $y_q(t)$, com $q \in \{0, \ldots, Q-1\}$;
2) De cada quadro $y_q(t)$, estimar o desvio padrão σ_q do ruído utilizando o estimador DATE [14];
3) Subtrair o desvio padrão de cada amostra do sinal corrompido para obter as amostras do sinal realçado, ou seja, $\tilde{y}_q(t) = \max\{y_q(t) - \sigma_q; 0\}$

Finalmente, os quadros são concatenados para formar o sinal de voz realçado $\hat{y}(t)$.

[1] Os códigos fonte dos métodos de realce EMDH e TASE estão disponíveis em http://lasp.ime.eb.br.



(a) (b) (c) (d) (e)

Fig. 1. Valores de INS obtidos de segmentos de 3 segundos dos ruídos acústicos: (a) avião, (b) balbúrdia, (c) fábrica, (d) trem, e (e) serra elétrica. As linhas tracejadas indicam o limiar de estacionariedade γ.

IV. Experimentos Realizados

Os experimentos de identificação de locutor são realizados com um subconjunto de 168 locutores da base de voz TIMIT [20]. As locuções possuem duração média de 3 segundos e taxa de amostragem de 16 kHz. Para cada locutor, oito locuções são concatenadas para gerar o modelo, e duas locuções são separadas para os testes. Para a formação das matrizes de atributos, foram utilizados 12 coeficientes MFCC e vetores pH com 6 componentes. Assim, vetores de atributos de 18 componentes são obtidos considerando janelas de 32 ms e 50% de sobreposição. Para os vetores pH, os expoentes de Hurst são estimados a partir das amostras de 2 janelas consecutivas. As matrizes resultantes são utilizadas para obter os modelos dos locutores com $M = 32$ componentes Gaussianas.

Cinco ruídos acústicos foram adicionados às locuções de teste considerando três valores de SNR: 10 dB, 15 dB e 20 dB. Eles foram selecionados das bases NOISEX-92 [21] (avião, balbúrdia e fábrica), Freesound.org[2] (trem) e Freesfx.co.uk[3] (serra elétrica). Neste trabalho, o índice de não-estacionariedade (INS - *index of nonstationarity*) [22] é utilizado para classificar os ruídos acústicos. Os valores de INS obtidos de segmentos de 3 s estão ilustrados na Fig. 1. A escala de tempo T_h/T indica a relação entre o tamanho da janela de tempo utilizada na análise espectral de tempo curto (T_h) e a duração total ($T = 3$ s) do ruído. Os valores de γ representam os limiares do teste de estacionariedade, com grau de confiança de 95%. Assim,

$$\text{INS} \begin{cases} \leq \gamma & \text{, ruído é estacionário;} \\ > \gamma & \text{, ruído é não-estacionário.} \end{cases} \quad (14)$$

Note a partir da Fig. 1 que o ruído avião é estacionário (E), já que INS $< \gamma$ para a maioria das escalas de tempo. Por possuir valores de INS no intervalo $\gamma <$ INS < 10, o ruído fábrica é aqui definido como moderadamente não-estacionário (MNE). Já os ruídos balbúrdia e trem apresentam o maior valor de INS no intervalo $[20, 40]$ e, por este motivo, são aqui denominados não-estacionários (NE). Finalmente, o ruído serra elétrica é classificado como altamente não-estacionário (ANE) visto que possui valor de INS > 90.

A Tabela I apresenta as taxas de acertos da tarefa de identificação de locutor com locuções de testes corrompidas pelos cinco ruídos acústicos. Os resultados estão ordenados segundo a não-estacionariedade dos ruídos. As duas primeiras colunas são referentes ao sistema de identificação com e sem o multitreinamento. O MT é implementado a partir de $R = 3$ sequências de ruídos Gaussianos com espectros coloridos: branco, rosa e marrom. Estas sequências são obtidas com o gerador proposto em [23] e são utilizadas para corromper as locuções de treinamento com SNR de 20 dB. Note que o método MT leva a um aumento nas taxas de acertos em 12 das 15 condições de ruídos. Por exemplo, a acurácia da identificação é incrementada em mais de 15 pontos percentuais (p.p.) para os ruídos trem e fábrica considerando SNR de 10 dB. As únicas situações onde o MT não melhora as taxas de acertos são para as fontes de ruído balbúrdia (15 dB e 20 dB) e trem (20 dB). Na média, o MT alcança os melhores resultados para todas as fontes de ruído.

TABELA I
Taxa de acertos (%) de identificação de locutor com diferentes ruídos acústicos.

Ruído	SNR	sem realce ou MT	MT	MT + UMMSE	MT + EMDH	MT + TASE
Serra elétrica (ANE)	20	86,9	87,2	85,1	92,3	**94,9**
	15	61,9	67,0	75,6	83,0	84,5
	10	23,5	30,4	51,5	**55,7**	**55,1**
Balbúrdia (NE)	20	97,0	89,6	88,7	94,0	94,3
	15	86,9	83,3	85,1	87,5	86,9
	10	55,1	68,8	68,5	67,9	66,1
Trem (NE)	20	92,6	90,8	86,0	91,4	**96,1**
	15	75,0	83,9	80,4	87,2	91,7
	10	49,7	65,5	69,6	69,3	69,3
Fábrica (MNE)	20	86,3	92,0	88,7	**95,2**	**94,9**
	15	66,4	81,0	85,7	92,9	89,0
	10	41,1	58,3	77,4	**84,8**	63,4
Avião (E)	20	48,5	63,4	77,1	59,8	**82,7**
	15	15,8	34,5	50,9	40,2	55,4
	10	2,4	14,9	15,2	**17,9**	**18,2**

A Tabela I também apresenta os resultados de identificação de locutor obtidos com a combinação do multitreinamento com o realce de sinais de voz. As três técnicas de realce, UMMSE, EMDH e TASE, são aplicadas considerando quadros de 32 ms de duração. Na fase de treinamento, o realce é aplicado sobre as locuções de voz corrompidas pelos três ruídos de espectro colorido, antes da extração da matriz de atributos. Os valores em destaque indicam as maiores taxas de acertos para SNR de 10 dB e 20 dB quando comparadas com a identificação com e sem o MT. Observe que, em geral, os maiores ganhos nas taxas de acertos são obtidos com a combinação do MT

[2]Disponível em http://www.freesound.org.
[3]Disponível em http://www.freesfx.co.uk.

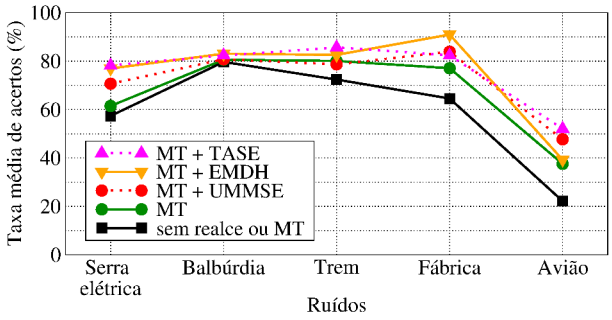


Fig. 2. Taxa média de acertos de identificação de locutor considerando as cinco fontes de ruído.

com as técnicas de realce no domínio do tempo. Esta melhora é obtida até mesmo para o ruído altamente não-estacionário serra elétrica. Para SNR de 10 dB, por exemplo, as propostas EMDH e TASE aumentam as taxas de acertos em mais de 20 p.p. quando comparadas com o MT apenas.

A combinação MT + TASE leva aos melhores resultados para quatro fontes acústicas considerando SNR de 20 dB. Os maiores incrementos na acurácia são obtidos para o ruído avião, onde o realce TASE leva a diferenças de 19.3 p.p. (SNR de 20 dB) e 20.9 p.p. (SNR de 15 dB) em comparação com o MT apenas. Já a combinação MT + EMDH consegue um ganho de 43.7 p.p. para o ruído Fábrica e SNR de 10 dB.

A Fig. 2 ilustra as taxas médias de acertos obtidas pelas diferentes configurações da identificação de locutor. Veja que a combinação MT + TASE obtém os melhores resultados para três fontes acústicas (serra elétrica, trem e avião), enquanto MT + EMDH atinge a maior acurácia para dois ruídos (balbúrdia e fábrica). Cabe ressaltar que mesmo para o ruído balbúrdia as técnicas temporais de realce conseguiram aumentar a taxa média de acertos quando comparadas ao treinamento em múltiplas condições. Esta fonte acústica é um grande desafio para as técnicas de realce, visto que as componentes do sinal e do ruído ocupam a mesma faixa de frequência.

V. Conclusão

Este artigo investigou um método combinado de realce de sinais de voz e multitreinamento para aumentar a robustez de sistemas de identificação de locutor. O multitreinamento foi implementado utilizando três ruídos acústicos de espectro colorido para corromper as locuções de treinamento. Para o realce, foram avaliadas três técnicas, sendo duas no domínio do tempo e uma no domínio da frequência. Os experimentos de identificação de locutor foram realizados com sinais de voz corrompidos por cinco ruídos acústicos. Estes ruídos foram classificados de acordo com o seu índice de não-estacionariedade. Os resultados demonstraram que o uso do realce no domínio do tempo combinado com o multitreinamento levou às melhores taxas de acertos da identificação. Esta combinação alcançou os melhores resultados mesmo para a fonte de ruído altamente não-estacionária. Na média, a maior acurácia foi obtida utilizando a técnica de realce TASE.

Referências

[1] 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.

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

[3] 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.

[4] 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.

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

[6] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, May 2012.

[7] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403–2418, November 2001.

[8] T. Hasan and M. Hasan, "Suppression of residual noise from speech signals using empirical mode decomposition," *IEEE Signal Processing Letters*, vol. 16, pp. 2–5, January 2009.

[9] 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.

[10] L. Zão, R. Coelho, and P. Flandrin, "Speech enhancement with EMD and Hurst-based mode selection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 899–911, May 2014.

[11] R. Tavares and R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Processing Letters*, vol. 23, pp. 6–10, January 2016.

[12] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and 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] E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, pp. 770–799, April 1951.

[14] D. Pastor and F.-X. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurences," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1545–1555, 2012.

[15] R. Sant'Ana, R. Coelho, and A. Alcaim, "Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 931–940, May 2006.

[16] A. Venturini, L. Zão, and R. Coelho, "On speech features fusion, α-integration gaussian modeling and multi-style training for noise robust speaker classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1951–1964, December 2014.

[17] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*, pp. 4266–4269, March 2010.

[18] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1996)*, vol. 32, pp. 629–632, December 1996.

[19] R. Coelho and L. Zão, "Empirical mode decomposition theory applied to speech enhancement," in *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering and Processing* (R. Coelho, V. Nascimento, R. Queiroz, J. Romano, and C. Cavalcante, eds.), Boca Raton, Florida: CRC Press, 2015.

[20] 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.

[21] 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, pp. 247–251, July 1993.

[22] 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.

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