Estimação de Frequência Fundamental de Sinais Acústicos Ruidosos com Aprendizado de Máquina

A. Queiroz e R. Coelho

***Resumo*— Este artigo apresenta uma proposta no domínio do tempo para aprimoramento das estimativas da frequência fundamental (F_0) do método HHT-Amp em sinais de voz ruidosos. Os quadros do sinal de voz são classificados em alta/baixa frequência por meio de uma Rede de Aprendizado DCNN (*Deep Convolutional Neural Network*), e os candidatos são extraídos de acordo com os mais prováveis tipos de erros de estimação. Por fim, uma função custo é definida como critério de seleção do novo valor da F_0. Os resultados dos experimentos mostraram uma superioridade da solução proposta DCNN+HHT-Amp nos diferentes cenários quando comparada aos métodos competitivos.**

***Palavras-Chave*— Aprimoramento da frequência fundamental, Rede Neural Convolucional Profunda.**

***Abstract*— This letter presents a proposal in the time domain to improve the estimates of the fundamental frequency (F_0) of the HHT-Amp method in noisy speech signals. The voiced frames are classified as high/low frequency by a Deep Convolutional Neural Network (DCNN), and candidates are extracted according to the most probable types of estimation errors. Finally, a cost function is defined as criterion for selecting the new value of F_0. The results of the experiments showed the superiority of the proposed solution DCNN+HHT-Amp in different scenarios when compared to competitive methods.**

***Keywords*— Fundamental Frequency Improvement, Deep Convolutional Neural Network.**

I. Introdução

Ruídos acústicos são efeitos naturalmente presentes em ambientes e cenários urbanos reais. Estes podem afetar características inerentes à voz como a frequência fundamental ou *pitch* dos sinais sonoros [1]. A estimação da F_0 de forma apurada possui grande relevância em diversas áreas do processamento de sinais, tais como, codificação, síntese, reconhecimento de voz ou locutor. Além disso, o estudo dos harmônicos em sinais de voz ruidosos tem sido explorado em soluções propostas para ganho de inteligibilidade [2][3][4][5].

Alguns métodos de estimação da F_0 no domínio do tempo propostos na literatura como, por exemplo, o ACF (*Auto-Correlation Function*) [6] e o YIN [7], são baseados na função autocorrelação. Por outro lado, as soluções SHR (*Subharmonic-to-Harmonic Ratio*) [8] e SWIPE (*Sawtooth Waveform Inspired Pitch Estimator*) [9] realizam uma abordagem no domínio espectral. Além destas técnicas, outros estimadores como o SFF (*Single Frequency Filtering*) [10] e HHT-Amp [11] foram propostos recentemente para atuação em sinais ruidosos. Estes apresentam resultados interessantes de acurácia, inclusive para baixos valores de SNR [12].

Estudos recentes exploram o aprendizado de máquina (*Machine Learning*) na estimação ou aprimoramento da *pitch* em ambientes ruidosos [13]. Algoritmos de classificação têm sido investigados, os quais utilizam-se de atributos espectrais (energia e harmônicos) do sinal [14][15] ou diretamente das sequências de amostras [16][17]. Em [18] é apresentada uma proposta para aprimoramento da acurácia das estimativas da F_0 pela utilização de uma Rede Neural DCNN. Esta solução adota estimativas iniciais de *pitch* obtidas pelo método SHR. A partir destes valores, candidatos a F_0 são extraídos de acordo com a classificação em alta ou baixa frequência do quadro do sinal de voz realizada pela DCNN. Por fim, a frequência fundamental aprimorada é selecionada dentre os candidados considerando critérios de seleção que buscam minimizar os valores de custo associados a essas estimativas [18].

O presente artigo propõe a utilização do método investigado em [18] aplicado no aprimoramento das estimativas da *pitch* obtidas pelo estimador HHT-Amp descrito em [11]. A acurácia da proposta DCNN+HHT-Amp é avaliada em cenários ruidosos, considerando-se as duas principais medidas de erro definidas na literatura: GE (*Gross Error*) e MAE (*Mean Absolute Error*). Métodos competitivos são adotados para análise destes resultados, tais como as soluções convencionais SHR [8], SFF [10], HHT-Amp [11], e também composições DCNN+SHR e DCNN+SFF. Duas bases de sinais de voz são utilizadas para o treinamento e os testes: a CSTR (*Centre of Speech Technology Research*) [19] e a base TIMIT [20]. Ambas disponibilizam os valores da F_0 utilizadas como referência na avaliação das soluções. Os sinais de voz são corrompidos por ruídos acústicos provenientes de seis fontes (Balbúrdia, Cafeteria, Trem, SSN (*Speech Shaped Noise*), Helicóptero e Volvo), com quatro valores de SNR: -10 dB, -5 dB, 0 dB e 5 dB. Para examinar o comportamento dos diferentes sinais ruidosos, são apresentadas as medidas do índice de não-estacionaridade (INS) [21]. Extensivos experimentos são realizados, cujos resultados demonstram uma maior acurácia da solução proposta DCNN+HHT-Amp, quando comparada com os outros métodos competitivos.

O restante do artigo está organizado da seguinte forma: A Seção II demonstra os efeitos causados pelos cenários ruidosos nos sinais de voz, por meio das suas medidas de não-estacionaridade. A Seção III descreve os métodos de estimação de *pitch* convencionais, propondo a utilização da DCNN nas estimativas do método HHT-Amp. Na Seção IV, as técnicas são avaliadas de acordo com a acurácia da estimação para diferentes cenários ruidosos. Finalmente, a Seção V conclui o trabalho.

A. Queiroz é doutorando do Programa de Pós-Graduação em Engenharia de Defesa do Instituto Militar de Engenharia (IME) e Bolsista da CAPES. O trabalho dos autores A. Queiroz e R. Coelho é desenvolvido no Laboratório de Processamento de Sinais Acústicos (LASP/IME) e parcialmente financiado pelo CNPq (308155/2019-0) e pela FAPERJ (203075/2016). E-mails: {anderson.queiroz,coelho}@ime.eb.br.

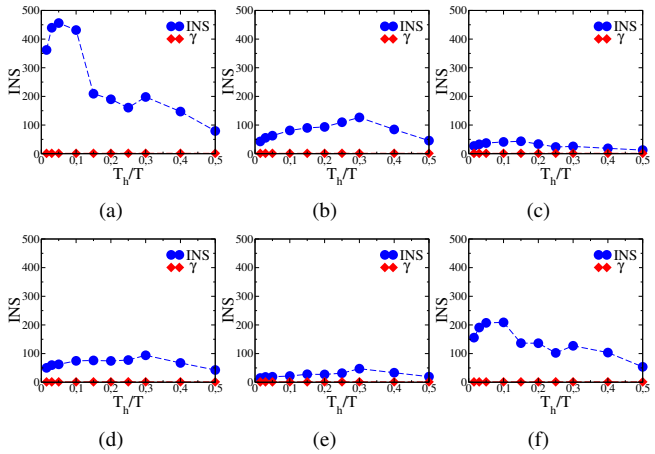

Fig. 1. Valores de INS para (a) Sinal de voz Limpo, e suas cinco versões ruidosas: (b) Balbúrdia, (c) Cafeteria, (d) Helicóptero, (e) SSN, (f) Volvo, com SNR = 0 dB.

II. Os Ruídos Acústicos e a Não-Estacionaridade em Sinais Sonoros

Em ambientes urbanos, os sinais de voz estão suscetíveis a interferências de ruídos acústicos. Estas interferências, provenientes de diversas fontes, podem afetar abruptamente a acurácia da estimação da F_0. Este desafio pode ser ainda maior quando o ruído é não-estacionário [11][12].

A medida INS (Índice de Não-Estacionaridade) [21] é adotada no presente trabalho para estudo dos efeitos dos cenários ruidosos nos sinais de voz. Esta medida é obtida a partir da comparação do sinal com referenciais estacionários chamados *surrogates*. O INS é obtido de acordo com a escala de observação T_h/T, que consiste na razão entre o tamanho da janela utilizada na análise espectral (T_h), e a duração total do sinal (T). Em [21], um limiar γ é definido para cada valor da janela T_h, considerando uma precisão de 95%. Este limiar é comparado com o valor de INS para avaliação da hipótese de estacionaridade, ou seja

$$\text{INS} \begin{cases} \leq \gamma, & \text{sinal é estacionário;} \\ > \gamma, & \text{sinal é não-estacionário.} \end{cases} \quad (1)$$

A Figura 1 retrata os valores de INS de um segmento de sinal de voz limpo extraído da base TIMIT [20], com duração de $T = 1,5$ s, e outras 5 versões ruidosas. Os ruídos selecionados para análise foram os ruídos SSN, da base DEMAND [22], Cafeteria e Helicóptero da base Freesound.org[1], e Balbúrdia e Volvo, coletados da base RSG-10 [23]. Segundo os resultados de INS, os sinais nas seis condições apresentam não-estacionaridade em todas as escalas temporais. Note na Figura 1(a) que o INS ressalta a natureza altamente não-estacionária do sinal de voz limpo. Esta característica da voz limpa é atenuada com as distorções por ruído, com valores de $\text{INS}_{\text{máx}}$ variando de 450 no sinal limpo, para menos de 100 nos sinais corrompidos pelos ruídos Cafeteria, Helicóptero e SSN. Estes efeitos podem afetar os componentes harmônicos do sinal (F_0 ou formantes), comprometendo a acurácia dos estimadores.

[1]Disponível em: www.freesound.org.

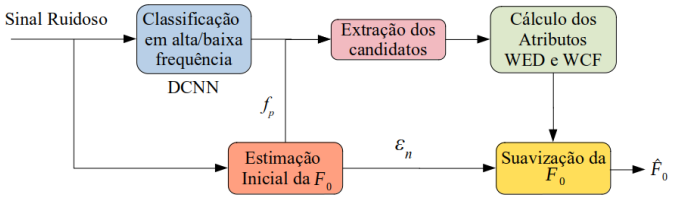

Fig. 2. Diagrama de blocos do algoritmo adaptado de [18].

III. Estimadores de F_0 em Sinais Ruidosos

A Figura 2 apresenta o diagrama esquemático do algoritmo proposto para aprimoramento das estimativas da *pitch*, o qual é descrito nesta Seção. Inicialmente, são enumerados os algoritmos de estimação da F_0 convencionais (sem aprendizado de máquina) adotados neste estudo. Então, o aprendizado de máquina DCNN é abordado, demonstrando como ocorre a classificação dos quadros do sinal em alta/baixa frequência. De acordo com esta classificação e a F_0 inicial (f_p), são apresentados os critérios de extração dos candidatos e seleção da *pitch* aprimorada. Por fim, destaca-se a proposta DCNN+HHT-Amp, como método capaz de elevar a acurácia das estimativas da F_0, inclusive de sinais em severas condições de ruídos.

A. Métodos de Estimação de F_0 Convencionais

A seguir, são enumerados os métodos de estimação convencionais adotados no desenvolvimento deste estudo.

1) ***SHR****:* Este estimador é definido em [8], e baseia-se na definição da medida denominada como Razão Sub-Harmônico-Harmônico. A relação entre a energia destes dois componentes, torna o algoritmo robusto na presença de sub-harmônicos. O SHR foi investigado em [18], e portanto é utilizado no presente trabalho como método comparativo.

2) ***SFF****:* Nesta solução, a frequência fundamental é extraída do sinal, usando uma técnica denominada filtragem de frequência única (SFF - *Single Frequency Filtering*), resultando em envelopes do sinal de voz filtrado [10]. Para extração da F_0, a função autocorrelação é aplicada ao envelope. Os picos da função autocorrelação localizados fora do intervalo $\tau_{min} \leq \tau_0 \leq \tau_{max}$ são desconsiderados. Desta forma é estabelecida uma faixa de frequências $[F_{min}, F_{max}]$, onde encontram-se os valores de F_0 dos sinais de voz [10].

3) ***HHT-Amp****:* Esta solução [11] baseia-se na aplicação da decomposição EEMD (*Ensamble Empirical Mode Decomposition*) [25]. Esta decomposição resulta em uma série de IMFs (*Intrinsic Mode Functions*), onde cada uma delas possui uma oscilação característica. O algoritmo EEMD é combinado com a transformada de Hilbert [26], resultando na transformada de Hilbert-Huang (HHT - *Hilbert Huang Transform*) [27]. Da HHT, derivam-se as amplitudes e frequências instantâneas das IMFs em função do tempo. Valores candidatos a F_0 são obtidos por meio do cálculo da função autocorrelação das amplitudes instantâneas das primeiras IMFs. Por fim, um critério definido em [11] é aplicado para selecionar o melhor candidato a *pitch*. Resultados expostos em [11] e [12] indicam que o HHT-Amp apresenta resultados interessantes, tanto para cenários ruidosos quanto em ambientes reverberantes-ruidosos, superando outros métodos de estimação em termos de acurácia.

TABELA I
CANDIDATOS DE ALTA/BAIXA FREQUÊNCIA DE ACORDO COM A ESTIMAÇÃO INICIAL f_p.

	Estimação Inicial f_p	Candidatos f_j para DCNN = Alta freq.	Candidatos f_j para DCNN = Baixa freq.
1	[50Hz, 66Hz]	$\{4f_p\}$	$\{f_p, 2f_p\}$
2	[66Hz, 100Hz]	$\{4f_p, 3f_p\}$	$\{f_p, 2f_p\}$
3	[100Hz, 133Hz]	$\{3f_p, 2f_p\}$	$\{f_p, 0{,}5f_p\}$
4	[133Hz, 200Hz]	$\{2f_p\}$	$\{f_p, 0{,}5f_p\}$
5	[200Hz, 400Hz]	$\{f_p\}$	$\{0{,}5f_p\}$
6	> 400Hz	$\{0{,}5f_p\}$	-

B. Métodos de Estimação da F_0 com Aprendizado de Máquina

O método de aprendizado considerado neste estudo é apresentado em [18], o qual consiste no treinamento de uma Rede Neural Convolucional Profunda (DCNN). A DCNN é implementada na classificação dos quadros de voz do sinal em Alta ($F_0 > 200$ Hz) ou Baixa ($F_0 \leq 200$ Hz) frequência. Para a entrada da Rede Neural, são extraídos quadros sonoros de 60 ms do sinal com taxa de amostragem de 16 kHz, ou seja, 960 amostras. A arquitetura da DCNN adotada em [18] baseia-se na VGGNet [28], sendo composta por seis camadas convolucionais, três camadas *Fully-Connected* (FC) e uma camada classificadora na saída (*Softmax*). Apesar da VGGNet ter sido desenvolvida para tarefas de classificação de imagens, esta arquitetura eleva a acurácia da Rede Neural para amostras unidimensionais, como é o caso dos sinais de voz.

Em [18], após a classificação dos sinais em alta/baixa frequência, o estimador SHR é implementado, para obter as estimativas iniciais (f_p). Desta forma, o próximo passo consiste na extração dos candidatos a *pitch* (f_j). A Tabela I mostra o critério de seleção dos candidatos. Note que para um sinal classificado como alta ou baixa frequência, os valores dos candidatos devem estar entre [200,400]Hz e [50,200]Hz, respectivamente.

Para auxiliar na seleção da F_0 aprimorada dentre os candidatos, dois atributos espectrais são calculados: WED (*Weighted Euclidean Deviation*) e WCF (*Weighted Comb Filtering*). O WED analisa as posições dos picos de frequência para detectar o valor da *pitch*. Os cinco primeiros picos de frequência de cada quadro do sinal são encontrados, compondo o vetor de picos:

$$\mathbf{P_n} = [P_{1,n}, P_{2,n}, P_{3,n}, P_{4,n}, P_{5,n}]. \quad (2)$$

Para o candidato a *pitch* $f_{j,n}$, em Hz, um vetor de candidatos é calculado como

$$V_{j,n} = [v_{j1,n}, v_{j2,n}, v_{j3,n}, v_{j4,n}, v_{j5,n}], \quad (3)$$

onde $v_{ji,n} = P_{i,n}/f_{j,n}$ ($1 \leq i \leq 5$). Para sinais sem presença de ruídos, onde os picos de frequência encontrados localizam-se nos múltiplos harmônicos da *pitch*, o vetor de candidatos é igual a $T = [1, 2, 3, 4, 5]$. Desta forma, o atributo WED é calculado pela distância entre T e o vetor de candidatos

$$d_{j,n} = \left\| (V_j - T) \odot U \right\|_2, \quad (4)$$

onde $\odot$ consiste na multiplicação ponto a ponto, e $U_i = 1/T_i$.

O atributo WCF também é obtido para cada candidato $f_{j,n}$, a partir de

$$y_{j,n} = \sum_f X_n(f)\, C(f/f_{j,n}), \quad (5)$$

onde $X_n(f)$ é a energia do quadro n do sinal de voz, e $C(f/f_{j,n}) = \left(\frac{1}{2} + \frac{1}{2}\cos(2\pi f/f_{j,n}) \exp(-f/f_{j,n})\right)$. Este atributo destaca os componentes próximos de $f_{j,n}, 2f_{j,n}, 3f_{j,n}, \cdots$, de modo que o valor máximo de $y_{j,n}$ ocorre quando $f_{j,n}$ está próximo do valor real da *pitch*.

A definição da estimativa aprimorada $\hat{F}_0$ ocorre por meio da minimização de uma função custo. Esta é calculada considerando os atributos encontrados anteriormente, a partir da equação:

$$\text{cost}_n = |\log f_{j,n} - \log f_{i,n+1}| + \frac{\lambda}{pr_n\left(\frac{y_{j,n}}{\alpha} + \frac{1}{d_{j,n}} + \varepsilon_n\right)}, \quad (6)$$

onde $|\log f_{j,n} - \log f_{i,n+1}|$ é uma distância entre candidatos de dois quadros sucessivos, definida para impor uma suavização na curva das estimativas da *pitch*. As variáveis λ e α são parâmetros de regularização, com valores de 1,4 e 1,7, respectivamente. O parâmetro pr_n é igual a maior probabilidade da camada de saída (*Softmax*) da DCNN. Por fim, $\varepsilon_n = 1$ se $f_{j,n} = f_p$ e igual a zero nos demais casos.

Em [18], o estimador SHR é adotado para obtenção das estimativas iniciais (f_p). Entretanto, este método tem sua acurácia comprometida na atuação em sinais de voz severamente corrompidos por ruídos. Desta forma, o presente estudo propõe a utilização do método HHT-Amp aliado ao algoritmo de aprendizado DCNN (DCNN+HHT-Amp). Este estimador mostrou-se apurado na estimação da F_0 de sinais ruidosos não-estacionários [11], inclusive para valores baixos de SNR [12]. Além do método proposto DCNN+HHT-Amp, o estimador SFF também é investigado neste trabalho (DCNN+SFF), para fins de análise comparativa.

IV. EXPERIMENTOS, RESULTADOS E DISCUSSÃO

Esta Seção apresenta os resultados dos experimentos (GE e MAE) da solução proposta DCNN+HHT-Amp em comparação com os métodos SHR, SFF, HHT-Amp, DCNN+SHR e DCNN+SFF. As bases de sinais de voz CSTR [19] e TIMIT [20] são adotadas na avaliação. Ambas as bases disponibilizam a F_0 de referência dos sinais. A CSTR é composta por 100 locuções (50 masculinas e 50 femininas), e a TIMIT possui 192 sinais de 24 locutores (16 masculinos e 8 femininos). Ambas as bases apresentam taxa de amostragem de 16 kHz. Os ruídos são adicionados aos sinais de voz, considerando seis fontes distintas: SSN da base DEMAND [22]; Balbúrdia e Volvo, da RSG-10 [23]; e Cafeteria, Trem e Helicóptero da base Freesound.org. Quatro valores de SNR são avaliados: -10 dB, -5 dB, 0 dB e 5 dB.

Para o treinamento da DCNN, os segmentos sonoros do sinal de voz são divididos em quadros de 60 ms sobrepostos, com deslocamento de 10 ms. No processo de aprendizado são adotados 30% dos quadros da base CSTR e 30% da base TIMIT. Desta forma, o conjunto dispõe de 197130 quadros sonoros, correspondentes aos sinais limpo e corrompidos com os ruídos citados anteriormente, com SNR de -10 dB e 0 dB.

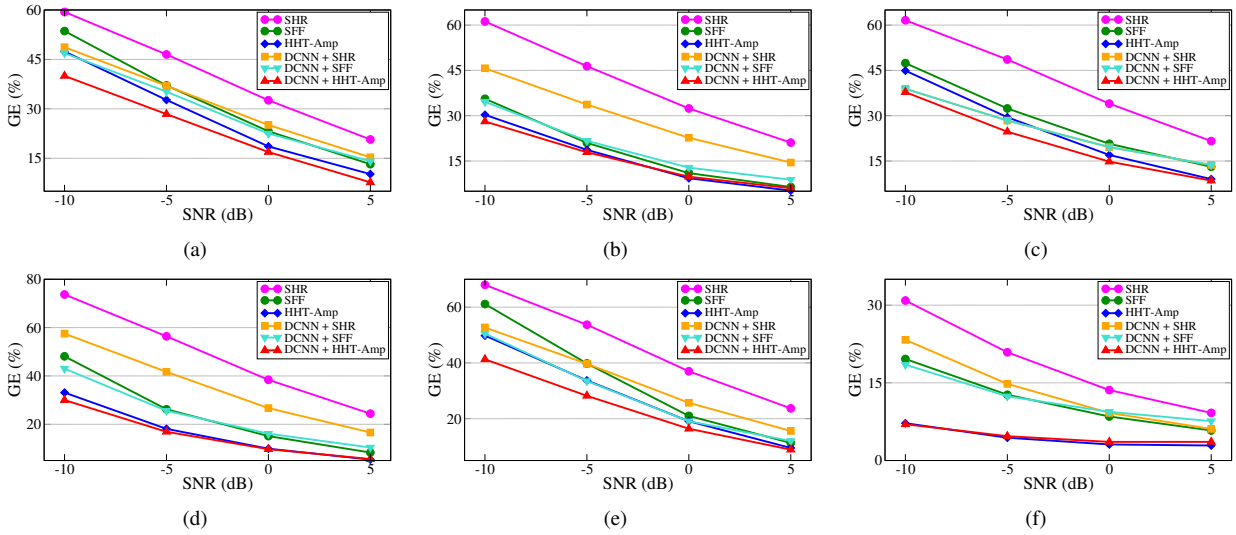

Fig. 3. Resultados de GE Médio das bases CSTR e TIMIT para ruídos (a) Balbúrdia ($\text{INS}_{\text{máx}} = 34,6$), (b) Trem ($\text{INS}_{\text{máx}} = 18,8$), (c) Cafeteria ($\text{INS}_{\text{máx}} = 11,7$), (d) Helicóptero ($\text{INS}_{\text{máx}} = 1,8$), (e) SSN ($\text{INS}_{\text{máx}} = 1,6$), (f) Volvo ($\text{INS}_{\text{máx}} = 0,9$), e quatro valores de SNR.

TABELA II
TEMPO MÉDIO DE PROCESSAMENTO NORMALIZADO.

SHR	SFF	HHT-Amp	DCNN+ SHR	DCNN+ SFF	DCNN+ HHT-Amp
0,02	0,37	0,87	0,20	0,53	1,00

A Tabela II apresenta a complexidade computacional referente ao tempo de processamento requerido por cada algoritmo avaliado para 512 amostras por quadro. Estes valores foram obtidos por uma máquina com processador Intel (R) Core (TM) i5-8400, com 8 GB de memória, cujos resultados são normalizados pelo tempo de execução da proposta DCNN+HHT-Amp. Note que os métodos HHT-Amp e DCNN+HHT-Amp apresentam maior tempo de processamento, visto que baseiam-se na decomposição EMD, a qual demanda um custo computacional relevante.

A. *Métricas de Avaliação dos Resultados*

As métricas de erro adotadas na comparação dos resultados incluem as medidas GE e MAE. A GE é aplicada na literatura pelos métodos comparativos [9][7][10][11]. O valor é obtido por $GE = P_{\text{erro}}/P \times 100$, onde P consiste no total de quadros sonoros e P_{erro} o número de quadros onde a estimativa $\hat{F}_0$ difere-se em mais de 20% da F_0 de referência.

A medida MAE possibilita uma maior percepção do erro, visto que indica uma distância absoluta (em Hz) entre a F_0 de referência e a estimada. O erro é definido por $MAE = \left(\sum_{i=1}^{n} \left|\hat{F}_0(i) - F_0(i)\right|\right)/n$, onde n denota a quantidade total de quadros sonoros, $\hat{F}_0(i)$ é a estimação e $F_0(i)$ a referência.

B. *Resultados de GE*

A Figura 3 apresenta a acurácia dos métodos competitivos, onde os resultados consistem nos valores médios de GE obtidos para os sinais ruidosos das bases CSTR e TIMIT. Note que a proposta DCNN+HHT-Amp supera os demais métodos nos diferentes cenários, inclusive nos valores negativos de SNR. Assim, esta solução demonstra-se interessante para as mais severas condições como, por exemplo, SNR = -10 dB. Nestes casos, o DCNN+HHT-Amp obteve os menores resultados de

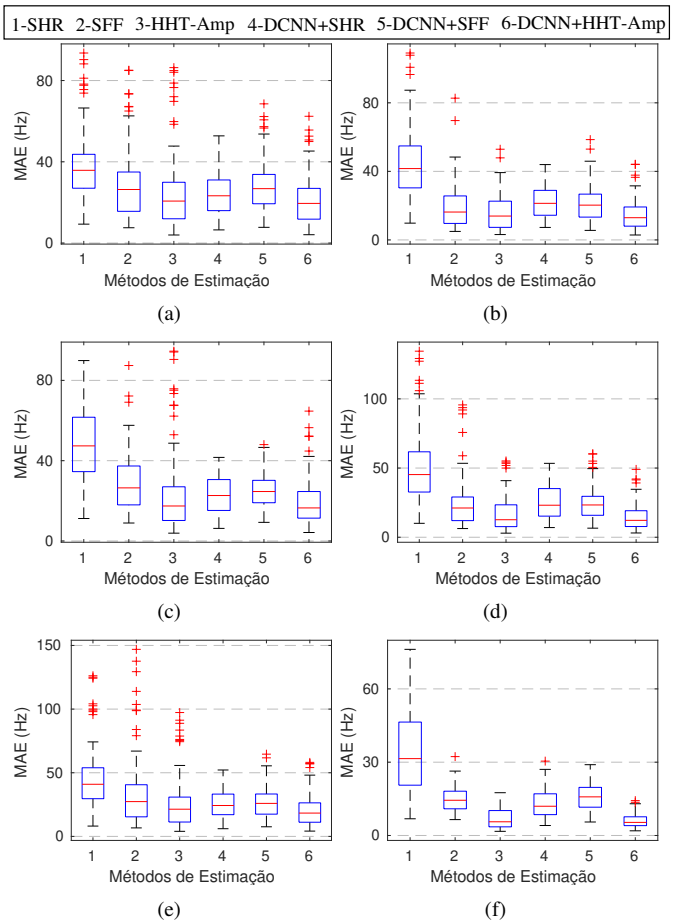

Fig. 4. Resultados de MAE para sinais de voz corrompidos por (a) Balbúrdia, (b) Trem, (c) Cafeteria, (d) Helicóptero, (e) SSN e (f) Volvo, considerando quatro valores de SNR.

GE. Observe ainda que o estimador convencional HHT-Amp mostra-se mais apurado que os métodos com aprendizado de máquina DCNN+SHR e DCNN+SFF.

O melhor aprimoramento alcançado pelo método proposto pode ser observado no ruído SSN, com SNR = -10 dB. Neste caso, o método HHT-Amp apresentou taxa de GE de 49,8%,

enquanto DCNN+HHT-Amp reduziu este valor para 41,3%. Além disso, os resultados de GE são altamente dependentes do grau de não-estacionaridade do ruído acústico. Considerando, por exemplo, SNR = -10 dB, o valor de GE da proposta varia de 40,0% com o ruído Balbúrdia ($\text{INS}_{\text{máx}} = 34,6$) para 30% com Helicóptero ($\text{INS}_{\text{máx}} = 1,8$), e apenas 7,0% para o ruído Volvo ($\text{INS}_{\text{máx}} = 0,9$).

C. Resultados de MAE

Na Figura 4, pode-se observar os resultados das medidas do erro médio absoluto MAE para os seis cenários ruidosos. Cada diagrama de caixa representa a distribuição dos resultados para os quatro valores de SNR (-10 dB, -5 dB, 0 dB e 5 dB). Note que assim como na medida GE, na MAE a solução DCNN+HHT-Amp atingiu os menores resultados em todos os cenários. Neste método, destaca-se uma redução na mediana de 1,2 Hz para o ruído não-estacionário Balbúrdia em comparação com o estimador HHT-Amp convencional. Por outro lado, o método comparativo DCNN+SFF não apresentou um aprimoramento relevante, e em alguns casos, como para o ruído Volvo, houve uma elevação dos índices de MAE (14,4 Hz para 15,8 Hz). Além disso, pode-se verificar que o método SHR atingiu os maiores valores de MAE, para todos os casos.

A partir dos resultados avaliados, verifica-se uma maior capacidade do estimador HHT-Amp de atuar em sinais de voz corrompidos por ruídos de diferentes graus de não-estacionaridade. Este fato reforça a ideia proposta neste trabalho, da implementação do método de aprendizado de máquina (DCNN) juntamente com o estimador HHT-Amp.

V. Conclusão

Este artigo apresentou a proposta DCNN+HHT-Amp para aprimoramento das estimativas da frequência fundamental de sinais de voz em cenários ruidosos. Nesta solução, uma DCNN classifica os quadros do sinal, e define candidatos a *pitch* de acordo com as estimativas do método HHT-Amp. Por fim, critérios são adotados para a seleção da F_0 aprimorada. Seis ruídos foram considerados na avaliação, com quatro valores de SNR. Os resultados de acurácia mostraram que a solução DCNN+HHT-Amp reduziu os valores de GE em até 17%, superando os demais métodos comparativos. Os resultados mais interessantes são observados em cenários severamente afetados por ruídos (SNR = -10 dB) de diferentes graus de não-estacionaridade, onde as taxas de erro obtidas foram menores que 42% em todos os casos.

Referências

[1] D. Ealey, H. Kelleher e D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," *Proceedings of the EUROSPEECH.*, pp. 437-440, 2001.

[2] Y. Lu e M. Cooke, "The contribuition of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication.*, v. 51, pp. 1253-1262, 2009.

[3] H. Hong, Z. Zhao, X. Wang, e Z. Tao, "Detection of dynamic structures of speech fundamental frequency in tonal languages," *IEEE Signal Processing Letters.*, v. 17, no. 10, pp. 843-846, Oct. 2010.

[4] J. Chen, H. Yang, e X. Wu, "The effect of F0 contour on the intelligibility of speech in the presence of interfering sounds for Mandarin Chinese," *The Journal of the Acoustical Society of America.*, v. 143, no. 2, pp. 864-877, 2008.

[5] L. Wang, D. Zheng e F. Chen, "Understanding low-pass-filtered Mandarim sentences: Effects of fundamental frequency contour and single-channel noise suppression," *Acoustical Society of America*, v. 143 no. 3, pp. 141-145, Mar. 2018.

[6] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech Signal Process.*, v. 25, pp. 24-33, Feb. 1977.

[7] A. de Cheveigné e H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, v. 111, no. 4, pp. 1917-1930, Apr. 2002.

[8] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, v. 1, pp. 333-336, 2002.

[9] A. Camacho e J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Amer.*, v. 124, no. 3, pp. 1638-1652, Sep. 2008.

[10] G. Aneeja and B. Yegnanarayana, "Extraction of fundamental frequency from degraded speech using temporal envelopes at high SNR frequencies," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, v. 25, no. 4, pp. 829-838, Apr. 2017.

[11] L. Zão and R. Coelho, "On the Estimation of Fundamental Frequency From Nonstationary Noisy Speech Signals Based on the Hilbert–Huang Transform," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 248–252, Feb. 2018.

[12] A. Queiroz, e R. Coelho, "Estudo de Métodos de Estimação de Frequência Fundamental em Sinais Reverberantes-Ruidosos," *XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT.*, 2020.

[13] T. Drugman, G. Huybrechts, V. Klimkov, e A. Moinet, "Traditional Machine Learning for Pitch Detection," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1745–1749, Nov. 2018.

[14] K. Han, e D. Wang, "Neural networks for supervised pitch tracking in noise," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 1502-1506, 2014.

[15] B. Liu, J. Tao, D. Zhang, e Y. Zheng, "A novel pitch extraction based on jointly trained deep BLSTM recurrent neural networks with bottleneck features," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 336-340, 2017.

[16] P. Verma, e R Schafer, "Frequency estimation from waveforms using multi-layered neural networks," *Proc. Interspeech*, pp. 2165-2169, 2016.

[17] J. Kim, J. Salamon, P. Li, e J. Bello, "CREPE: A convolutional representation for pitch estimation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 161-165, 2018.

[18] M. Khadem-hosseini, S. Ghaemmaghami, A. Abtahi, S. Gazor, e F. Marvasti, "Error Correction in Pitch Detection Using a Deep Learning Based Classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*, v. 28, pp. 990-999, Mar. 2020.

[19] P. C. Bagshaw, S. M. Hiller e M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," *Proc. EUROSPEECH?93.*, pp. 1003-1006, Sep. 1993.

[20] S. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *Philadelphia, PA, USA: NASA STI/Recon, Tech. Rep. N*, vol. 24, 1993.

[21] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, e J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, v. 58, no. 7, pp. 3459-3470, Jul. 2010.

[22] J. Thiemann, N. Ito, and E. Vincent, "Demand: A collection of multichannel recordings of acoustic noise in diverse environments," *Proc. Meetings Acoust.*, 2013.

[23] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise database," TNO Inst. Perception, Soesterberg, The Netherlands, Tech. Rep. IZF 3, 1988.

[24] R. Tavares e R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Processing Letters*, v. 23, pp. 6-10, Jan. 2016.

[25] Z. Wu e N. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, v. 1, no. 1, pp. 1-41, 2009.

[26] Huang, et al., " The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond.*, pp. 903-995, 1998.

[27] H. Huang e J. Pan, " Speech pitch determination based on Hilbert-Huang transform," *Signal Process.*, v. 86, no. 4 pp. 792-803, 2006.

[28] K. Simonyan, e A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. ICLR.*, 2015.