Frequência Fundamental de Sinais Reverberantes e Ruidosos com Classificação de Atributos Harmônicos

A. Queiroz e R. Coelho

***Resumo*— Este artigo apresenta uma análise da classificação dos atributos harmônicos na acurácia da estimação da frequência fundamental (F0) de sinais reverberantes e ruidosos. Duas soluções (DCNN-BPS e FSFFE) propostas inicialmente para aprimoramento da acurácia da estimação da F0 em sinais ruidosos são avaliadas neste trabalho para diferentes condições de reverberação e ruído. Os métodos de estimação SWIPE, SHR e HHT-Amp e duas medidas de erro (GE e MAE) são utilizadas na avaliação, onde os menores valores indicam melhor acurácia. Os resultados dos experimentos mostraram uma superioridade da solução FSFFE+HHT-Amp nos diferentes cenários quando comparada aos métodos competitivos.**

***Palavras-Chave*— Classificação de atributos harmônicos, Sinais reverberantes e ruidosos, Estimação da F0.**

***Abstract*— This letter presents an analysis of the classification of harmonic features in the accuracy of fundamental frequency (F0) estimation of noisy reverberant speech signals. Two solutions (DCNN-BPS and FSFFE) initially proposed to improve the F0 estimation accuracy in noisy signals are evaluated in this work for different reverberation and noise conditions. The SWIPE, SHR and HHT-Amp estimation methods and two error measures (GE and MAE) are used in the evaluation, where lower values indicate better accuracy. The results of the experiments showed a superiority of the FSFFE+HHT-Amp solution in the different scenarios when compared to the competitive methods.**

***Keywords*— Classification of harmonic features, Noisy reverberant signals, F0 Estimation.**

I. Introdução

Ruído e reverberação são efeitos acústicos comumente presentes em ambientes e cenários urbanos. Estes podem afetar os sinais de voz alterando suas características temporais e espectrais. Tais alterações podem ser notadas nos seus componentes harmônicos como a frequência fundamental (F0) ou *pitch* [1]. A estimação da F0 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 de sinais de voz reverberantes e ruidosos pode ser explorado em soluções para aprimoramento da inteligibilidade sonora [2][3][4].

Diferentes métodos de estimação da F0 de sinais sonoros ou harmônicos foram propostos na literatura com atuação tanto no domínio do tempo como no domínio espectral. Por exemplo, o ACF (*Auto-Correlation Function*) [8] e YIN [9], caracterizam-se por uma abordagem temporal baseada na função autocorrelação. Em contrapartida, os métodos SHR (*Subharmonic-to-Harmonic Ratio*) [10] e SWIPE (*Sawtooth Waveform Inspired Pitch Estimator*) [11] atuam no domínio da frequência. Além destas técnicas, outros estimadores como o SFF (*Single Frequency Filtering*) [12] e HHT-Amp [13] foram propostos para atuação em sinais ruidosos.

Recentemente, soluções na literatura buscam melhorar a acurácia das estimações da F0 em ambientes ruidosos [15][16]. Estas estratégias realizam uma classificação em alta/baixa frequência (*low/high pitch*) nos quadros harmônicos do sinal corrompido. Na técnica DCNN-BPS (*DCNN-Based Pitch Separation*) [15] novos candidatos a estimativa da F0 são gerados de acordo com esta classificação. Por outro lado, no método FSFFE (*Frequency Separation for Fundamental Frequency Estimation*) [16], a separação auxilia na correção dos candidatos extraídos de um estimador convencional, como por exemplo SHR e HHT-Amp.

O presente artigo propõe a investigação dos métodos FSFFE [16] e DCNN-BPS [15] para sinais reverberantes e ruidosos. As técnicas de estimação SWIPE, SHR e HHT-Amp também são consideradas na avaliação, que em composição com os métodos de classificação *low/high pitch* totalizam nove métodos comparativos. Os resultados são examinados para cenários reverberantes e ruidosos, considerando-se as duas principais medidas de erro de estimação presentes na literatura: GE (*Gross Error*) e MAE (*Mean Absolute Error*). A base de sinais de voz CSTR (*Centre of Speech Technology Research*) [17] é adotada nos experimentos. Os sinais são reverberados com a resposta ao impulso (RIR - *Room Impulse Response*) de duas salas reais: LASP2 da base LASP_RIR[1] e *Stairway* da base AIR [18]. Ademais, foram selecionados quatro ruídos acústicos (Balbúrdia, Cafeteria, Helicóptero e Trânsito) com valores de razão sinal-ruído (SNR - *Signal-to-Noise Ratio*) de -10 dB, -5 dB, 0 dB e 5 dB.

As principais contribuições deste trabalho são:

- Demonstrar o impacto da reverberação e ruído acústico na acurácia da estimação da F0;
- Analisar a classificação dos atributos harmônicos (*low/high pitch*) na acurácia da estimação da F0 de sinais reverberantes e ruidosos.

O restante do artigo está organizado da seguinte forma: A Seção II demonstra os efeitos causados pela reverberação e ruído na estimação da F0 de sinais de voz. A Seção III descreve os métodos de separação dos quadros do sinal em *low/high pitch* para aprimoramento da acurácia das estimativas da F0. Na Seção IV, os métodos comparativos são avaliados para cenários reverberantes e ruidosos, por meio das medidas GE e MAE. Por fim, 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 (305488/2022-8) e pela FAPERJ (200518/2023). E-mails: {anderson.queiroz, coelho}@ime.eb.br.

[1]Disponível em lasp.ime.eb.br

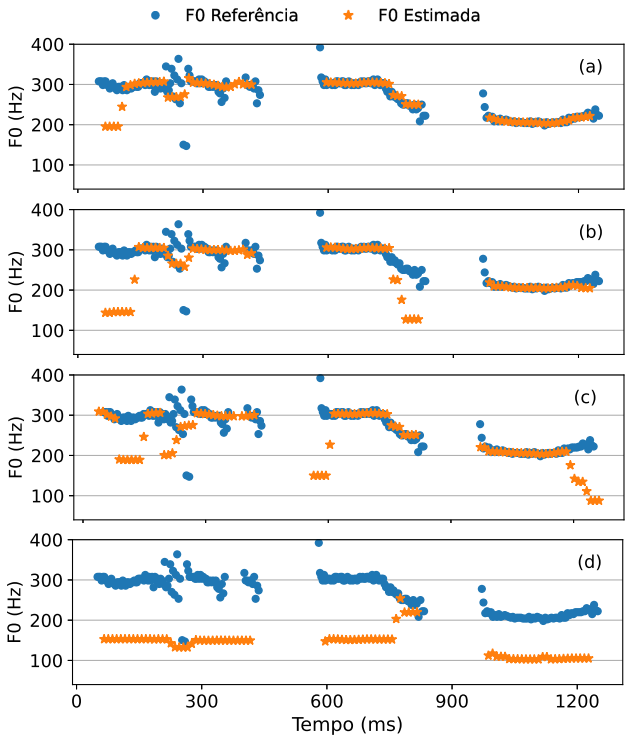


Fig. 1. Estimativa da Frequência Fundamental de um sinal de voz limpo (a), sinal reverberado com a RIR da sala LASP2 (RT_{60} = 0,79 s) (b), sinal ruidoso com ruído Balbúrdia e SNR = 0 dB (c), e sinal de voz reverberante e ruidoso com a sala LASP2 e Balbúrdia (SNR=0 dB) (d).

II. Efeitos Reverberantes e Ruidosos e o Impacto na Estimação da F0

Um sinal reverberado pode ser definido pela convolução do sinal de voz com a resposta ao impulso $h(t)$ de uma sala. A RIR é tipicamente caracterizada pelo tempo de reverberação RT_{60} e pela razão entre os sinais direto e reverberado (*Direct-to-Reverberant Ratio* - DRR). Estes parâmetros descrevem a duração da reverberação até uma redução de 60 dB na sua energia e a intensidade relativa do sinal direto, respectivamente. Além disso, os ruídos acústicos são interferências comumente presentes em ambientes urbanos (abertos ou fechados). Esses ruídos são geralmente definidos como um efeito aditivo ao sinal de voz. Deste modo, um sinal reverberante e ruidoso $x(t)$ pode ser descrito por $x(t) = s(t) * h(t) + w(t)$, onde $s(t)$ é a voz, e $w(t)$ o ruído acústico.

A Figura I ilustra as estimativas da frequência fundamental de um sinal de voz para quatro cenários: sinal limpo (a), sinal reverberante (b), sinal ruidoso (c), e reverberante e ruidoso (d). As estimativas da F0 são aplicadas nos quadros sonoros do sinal com o método HHT-Amp [13] a cada 10 ms e comparadas com a referência fornecida pela base. Note que a curva da F0 para o sinal limpo converge com a referência, apresentando boa acurácia. O sinal reverberante com a RIR da sala LASP2 apresenta uma distância da referência entre algumas estimativas assim como o sinal ruidoso. Esta diferença aumenta consideravelmente nos sinais reverberantes e ruidosos. Observe na Figura I (d) que neste cenário grande parte das estimativas não converge configurando assim, erros de estimação.

TABELA I
Valores de GE Médio para Diferentes Cenários com a Sala LASP2 e Ruído Balbúrdia

	SHR				HHT-Amp			
SNR (dB)	-5	0	5	Méd.	-5	0	5	Méd.
Limpo			9,4				5,0	
Reverberante			41,6				23,2	
Ruidoso	51,4	36,3	23,8	37,2	41,9	27,7	16,8	28,8
Rev. e Ruidoso	65,6	54,5	44,0	54,7	52,8	40,8	32,0	41,9

Para analisar o impacto dos efeitos reverberantes e ruidosos na acurácia da estimação da F0 aplica-se a medida GE (*Gross Error rate*). Esta medida é comumente adotada na literatura [11][9][12][13], e definida por

$$GE = P_{\text{erro}}/P \times 100, \quad (1)$$

onde P consiste no número total de quadros sonoros e P_{erro} o número de quadros onde a estimativa ($\hat{F}0$) difere-se em mais de 20% da $F0$ de referência.

A Tabela I apresenta os valores médios de GE da F0 estimada com os métodos SHR [10] e HHT-Amp [13]. Neste estudo adota-se as 100 locuções da base CSTR (50 masculinas e 50 femininas) com taxa de amostragem de 16 kHz [17]. Estes sinais são reverberados com a RIR da sala LASP2 e o ruído Balbúrdia com três valores de SNR: -5 dB, 0 dB e 5 dB. Observe que o GE médio dos sinais limpos está abaixo de 10% para ambos os métodos de estimação, ou seja, as estimativas da F0 apresentam boa acurácia. Quando o sinal limpo é reverberado, o valor de GE aumenta de 9,4% para 41,6% no SHR e de 5,0% para 23,2% para HHT-Amp. Da mesma forma, sinais ruidosos possuem valores de erro de estimação mais elevados que os limpos, com GE aumentando de acordo com a redução do valor de SNR. Note ainda que os ambientes reverberantes e ruidosos afetam ainda mais a acurácia das estimativas, compondo assim os cenários mais desafiadores para atuação dos métodos de detecção da F0. Para estes sinais, o GE médio atingido no pior caso é de 65,6% e 52,8% para o estimador SHR e HHT-Amp, respectivamente.

III. Soluções para Classificação dos Atributos Harmônicos em Baixa/Alta Pitch

Esta Seção descreve sucintamente os métodos DCNN-BPS [15] e FSFFE [16] para classificação dos quadros sonoros (harmônicos) do sinal em baixa ou alta *pitch*.

A. Solução FSFFE

O método FSFFE [16] foi proposto para separar os quadros harmônicos do sinal em baixa/alta *pitch*, aprimorando a acurácia da estimativa da F0 de acordo com esta separação. Esta estratégia é realizada em quatro passos principais:

1) O primeiro passo consiste em realizar uma decomposição tempo-frequência do sinal corrompido $x(t)$ com EEMD (*Ensamble Empirical Mode Decomposition*) [19][20]. O resultado da decomposição apresenta uma série de IMFs (*Intrinsic Mode Functions*) e um residual $r(t)$, de modo que $x(t) = \sum_{k=1}^{K} \text{IMF}_k(t) + r(t)$, onde cada IMF apresenta uma oscilação característica.

TABELA II
RESULTADOS DE ERRO (%) DE CLASSIFICAÇÃO EM BAIXA/ALTA FREQUÊNCIA

Ambiente		SNR(dB)	Balbúrdia				Cafeteria				Helicóptero				Trânsito				Méd.
			-10	-5	0	5	-10	-5	0	5	-10	-5	0	5	-10	-5	0	5	
Ruidoso		DCNN-BPS	37,1	30,3	24,3	19,5	36,6	29,3	22,3	18,6	31,0	24,0	18,7	17,4	25,0	18,5	15,5	16,0	24,0
		FSFFE	**25,1**	**14,5**	**7,7**	**5,0**	**25,0**	**13,8**	**7,8**	**4,9**	**10,3**	**4,9**	**4,0**	**3,3**	**4,4**	**3,5**	**2,9**	**2,8**	**8,7**
Rev. e Ruidoso	LASP2	DCNN-BPS	37,5	31,8	26,4	23,3	38,2	31,7	25,6	22,2	29,0	24,3	21,9	21,1	28,5	23,4	21,5	20,8	26,7
		FSFFE	**35,4**	**21,2**	**15,4**	**8,1**	**32,3**	**20,8**	**11,0**	**8,2**	**15,9**	**9,3**	**7,2**	**4,8**	**7,2**	**5,7**	**4,2**	**4,3**	**13,2**
	Stairway	DCNN-BPS	43,0	38,3	32,7	28,4	43,1	37,5	31,9	26,8	36,4	31,0	27,0	24,9	33,4	27,7	24,9	24,0	31,9
		FSFFE	**37,3**	**25,3**	**13,7**	**7,5**	**34,3**	**22,1**	**14,2**	**8,9**	**19,3**	**10,1**	**7,6**	**5,7**	**8,7**	**6,2**	**4,6**	**4,1**	**14,4**

2) O método PEFAC (*Pitch Estimation Filter with Amplitude Compression*) [21] é utilizado para estimar a F0 dos quadros das primeiras quatro IMFs. Assim, considerando $\hat{F}0_{k,q}$ o valor da F0 para o quadro q da $\text{IMF}_k\,(t)$, o vetor $\hat{F}0_q$ é composto por

$$\hat{F}0_q = [\hat{F}0_{1,q}, \hat{F}0_{2,q}, \cdots \hat{F}0_{K,q}], \tag{2}$$

onde $K = 4$. Conforme demonstrado em [16], adotou-se PEFAC por apresentar maior convergência com a referência da F0 nas IMFs em comparação com outros estimadores.

3) Uma diferença normalizada $\delta^q_{\hat{F}0}$ entre as estimativas $\hat{F}0_q$ das IMFs é computada para os sucessivos quadros, de modo a evitar discrepâncias nas estimações. Em seguida, uma média $\bar{F}0_q$ é obtida para as duas IMFs com valores da F0 mais aproximados, ou seja, com menores valores de $\delta^q_{\hat{F}0}$. Assim, o critério de classificação do quadro é definido por

$$\begin{cases} \bar{F}0_q \leq \gamma, & \text{quadro de baixa-frequência;} \\ \bar{F}0_q > \gamma, & \text{quadro de alta-frequência,} \end{cases} \tag{3}$$

onde γ = 200 Hz, uma vez que a variação da F0 dos sinais de voz comumente está compreendida entre 50-200 Hz para homens e 120-350 Hz para mulheres [22].

4) Os candidatos a F0 extraídos de um estimador convencional são corrigidos de acordo com a classificação dos quadros do sinal. Para este ajuste são considerados os principais tipos de erros de (*halving*, *doubling*) onde a estimação resulta em múltiplos da F0 verdadeira [23]. Por fim, os candidatos retornam ao método de estimação, resultando em uma F0 com maior acurácia.

B. Separação DCNN-BPS

Esta técnica [15] consiste em treinar uma rede neural convolucional profunda (DCNN - *Deep Convolutional Neural Network*) para classificar os quadros sonoros do sinal em baixa ($F0 \leq 200$ Hz) ou alta ($F0 > 200$ Hz) *pitch*. A arquitetura da DCNN adotada baseia-se na VGGNet (*Visual Geometry Group Network*) [26], com seis camadas convolucionais, três camadas FC (*Fully-Connected*) e uma camada de saída (*Softmax*). A entrada da DCNN é composta por segmentos de 60 ms extraídos diretamente dos quadros sonoros do sinal, com taxa de amostragem de 16 kHz, ou 960 amostras.

De acordo com a classificação do quadro em baixa/alta frequência, novos candidatos a F0 (f_j) são obtidos, baseados na estimativa inicial da F0 (f_p) obtida por um método convencional. Dois atributos espectrais são adotados para auxiliar na seleção da F0 com acurácia aprimorada dentre os candidatos: WED (*Weighted Euclidean Deviation* - $d_{j,q}$) e WCF (*Weighted Comb Filtering* - $y_{j,q}$). Por fim, uma função custo é definida como

$$\text{cost}_q = |\log f_{j,q} - \log f_{i,q+1}| + \frac{\lambda}{pr_q(\frac{y_{j,q}}{\alpha} + \frac{1}{d_{j,q}} + \varepsilon_q)}, \tag{4}$$

onde $|\log f_{j,q} - \log f_{i,q+1}|$ é um fator de suavização da F0, $\lambda > 0$ e α são parâmetros de regularização, pr_q é a probabilidade de baixa/alta frequência da camada de saída *Softmax* da DCNN, e $\varepsilon_q = 1$ se $f_j = f_p$, ou zero caso contrário. O menor valor obtido pela função indica a estimativa mais apurada da F0 dentre os candidatos.

IV. EXPERIMENTOS, RESULTADOS E DISCUSSÃO

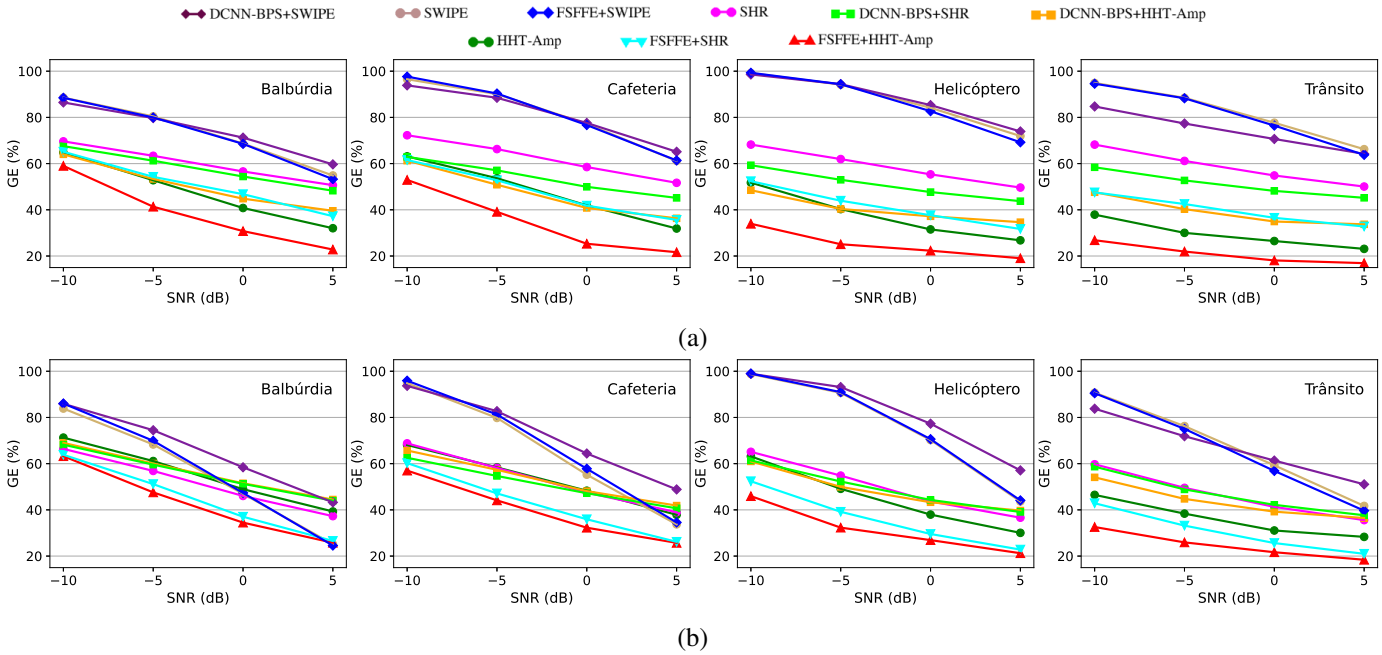
Esta Seção apresenta os resultados dos experimentos que incluem os erros de separação dos quadros dos sinais em baixa/alta frequência com a solução FSFFE em comparação com o método DCNN-BPS. Em seguida, as métricas de erro GE e MAE são adotadas na avaliação dos erros de estimação da F0 com SWIPE, SHR e HHT-Amp para sinais reverberantes e ruidosos. As composições destes estimadores com FSFFE e DCNN-BPS também são avaliadas em termos de redução das medidas de erro.

A base de sinais de voz CSTR [17] é adotada na avaliação dos resultados, a qual é composta por 100 locuções. A F0 de referência dos sinais da base é disponibilizada por meio de um laringógrafo. Duas salas foram selecionadas para representar os efeitos da reverberação, com Respostas ao Impulso obtidas de ambientes urbanos reais: a sala LASP2 da base LASP_RIR e *Stairway* da base AIR [18]. Estas salas apresentam valores de RT_{60} e DRR de {0,79s;1,05s} e {-4,37;-5,28}, respectivamente. Finalmente, os sinais de voz reverberados são corrompidos pelos ruídos Balbúrdia da base RSG-10 [25]; Cafeteria, Helicóptero e Trânsito da base Freesound.org[2].

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 um subconjunto da base TIMIT [24]. Desta forma, o conjunto dispõe de 300000 quadros sonoros, correspondentes aos sinais limpo, e reverberantes e ruidosos com as salas LASP2 e *Stairway* e os quatro ruídos com SNR de -10 dB e 0 dB. Os demais parâmetros da DCNN são os mesmos que o proposto em [15].

Os resultados de erro da separação dos quadros em *low/high pitch* com os métodos competitivos FSFFE e DCNN-BPS podem ser verificados na Tabela II. Note que os sinais reverberantes e ruidosos apresentam aumento nos erros em relação

[2]Disponível em www.freesound.org.



(a)

(b)

Fig. 2. Resultados de GE dos sinais reverberantes e ruidosos para sala LASP2 (a) e *Stairway* (b), considerando os ruídos acústicos Balbúrdia, Cafeteria, Helicóptero e Trânsito, e quatro valores de SNR: -10 dB, -5 dB, 0 dB e 5 dB.

TABELA III
TEMPO MÉDIO DE PROCESSAMENTO NORMALIZADO.

Estimador da F0			DCNN-BPS			FSFFE		
SHR	SWIPE	HHT-Amp	SHR	SWIPE	HHT-Amp	SHR	SWIPE	HHT-Amp
0,03	0,03	0,93	0,22	0,20	1,07	0,94	0,95	1,00

aos ruidosos. Por exemplo, no sinal ruidoso com Balbúrdia e SNR de -10 dB o erro médio de separação é de 37,1% e 25,1% para DCNN-BPS e FSFFE, respectivamente. Já no cenário reverberante e ruidoso com a sala *Stairway* e mesma condição de ruído, os respectivos erros passam para 43,0% e 37,3%. Apesar disso, observe que a solução FSFFE supera a técnica DCNN-BPS em todos os cenários. Para as salas LASP2 e *Stairway* o erro médio total do FSFFE foi de 13,2% e 14,4%, valores cerca de 50% menores quando comparado com o método competitivo.

A Tabela III mostra 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 solução FSFFE+HHT-Amp. Vale ressaltar que nesta avaliação não foi incluído o tempo demandado para treinamento da rede neural da técnica DCNN-BPS.

A. *Resultados de GE e MAE*

A Figura 2 apresenta os valores médios de GE para os sinais reverberantes e ruidosos da base CSTR. Observe que a composição FSFFE+HHT-Amp apresenta as menores taxas de erro, ou seja, atinge melhor acurácia em comparação com os métodos competitivos. Além disso, vale ressaltar que a solução FSFFE demonstra interessante aprimoramento na acurácia das estimativas para SHR e HHT-Amp, mesmo assumindo os erros na separação em baixa/alta frequência. Por exemplo, o estimador HHT-Amp obtém valor de GE de 51,8% para a sala LASP2 (Figura 2(a)) e ruído Helicóptero com SNR -10 dB. Neste mesmo cenário, o GE para a composição FSFFE+HHT-Amp reduz para 34,0%, ou seja, uma redução de 17,8 p.p. (pontos percentuais) no erro de estimação. Ainda para a sala LASP2, a separação DCNN-BPS aprimora a acurácia das estimativas da F0 para o estimador SHR, mais ainda assim é superado por FSFFE+SHR.

A Figura 2(b) ilustra as curvas com os resultados de GE para a sala *Stairway* com as mesmas condições de ruídos. Note que para a mesma condição de ruído avaliada anteriormente (Helicóptero SNR -10 dB) ocorre uma redução de 17,3 p.p. no valor de GE do estimador HHT-Amp para FSFFE+HHT-Amp. Esta redução é pouco menor que a obtida para sala LASP2, uma vez que o erro de separação dos métodos aumenta para a sala *Stairway*. O cenário reverberante e ruidoso mais desafiador dos experimentos é composto pela sala *Stairway* com ruído Balbúrdia e SNR -10 dB. Nesta condição, todos os métodos competitivos atingem GE superior a 60%. Apesar disso, a solução FSFFE aprimora a acurácia da F0 dos estimadores, sendo que a composição FSFFE+HHT-Amp supera novamente os demais métodos.

A medida MAE possibilita uma maior percepção do erro, visto que indica uma distância absoluta (em Hz) entre a F0 de referência e a estimada. O erro é definido por

$$MAE = \left(\sum_{i=1}^{n} \left| \hat{F0}(i) - F0(i) \right| \right) / n, \qquad (5)$$

onde n denota a quantidade total de quadros sonoros, $\hat{F0}(i)$ é a estimação e $F0(i)$ a referência.

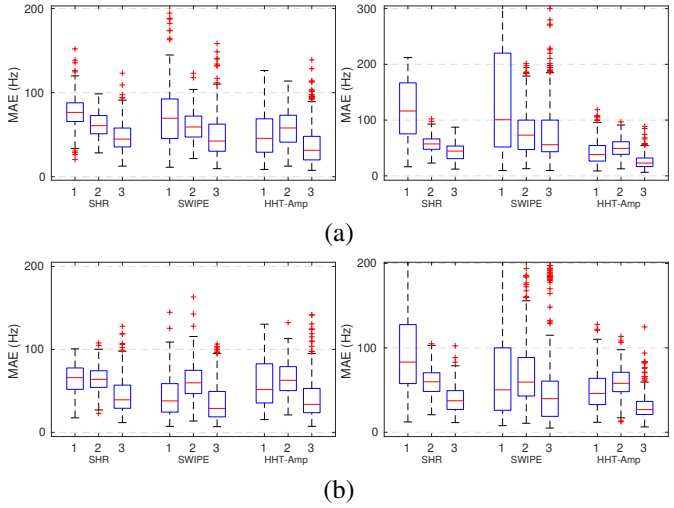


Fig. 3. MAE para sala LASP2 (a) e *Stairway* (b) para os ruídos Balbúrdia (esquerda) e Trânsito (direita) onde: Caso 1 - Estimador original; Caso 2 - DCNN-BPS+Estimador; Caso 3 - FSFFE+Estimador.

Na Figura 3 pode-se observar os resultados das medidas do erro médio absoluto MAE para as duas salas de reverberação e os ruídos Balbúrdia e Trânsito. 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 para todos os métodos de estimação da F0, a solução FSFFE é capaz de reduzir os valores de erro, superando a separação DCNN-BPS. Assim como na medida GE, a composição FSFFE+HHT-Amp apresenta os menores valores de MAE que os métodos comparativos. Neste método, destaca-se uma redução na mediana de 14,9 Hz para a sala *Stairway* com ruído Trânsito.

V. Conclusão

Este artigo apresentou um estudo dos efeitos reverberantes e ruidosos na acurácia das estimativas da frequência fundamental de sinais de voz. Além disso, avaliou-se a importância das separações *low/high pitch* DCNN-BPS e FSFFE nestes ambientes adversos, de modo a aprimorar a acurácia da F0 dos métodos de estimação SWIPE, SHR e HHT-Amp. Extensivos experimentos foram conduzidos utilizando duas salas de reverberação e quatro ruídos acústicos. As medidas GE e MAE foram adotadas na análise dos erros de estimação dos métodos competitivos. Os resultados de acurácia mostraram que a solução FSFFE+HHT-Amp superou os demais métodos comparativos, com menores valores de GE. Esta composição obteve resultados interessantes em cenários severamente impactados pelos efeitos reverberantes (*Stairway*) e ruidosos (SNR = -10 dB). Nesta condição, observa-se uma redução média de GE de 28,1% para todos os ruídos.

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] A. Queiroz e R. Coelho, "F0-Based Gammatone Filtering for Intelligibility Gain of Acoustic Noisy Signals," *IEEE Signal Processing Letters*, v. 28, pp. 1225–1229, 2021.

[4] A. Queiroz, e R. Coelho, "Métodos de Mitigação de Efeitos Reverberantes e Ruidosos com Ganho de Inteligibilidade e Qualidade," *XL Simpósio Brasileiro de Telecomunicações e Processamento de Sinais-SBrT.*, 2022.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] L. Zão and R. Coelho, "On the Estimation of Fundamental Frequency From Nonstationary Noisy Speech Signals Based on the Hilbert–Huang Transform," *IEEE Signal Process. Letters*, vol. 25, pp. 248–252, 2018.

[14] 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.

[15] 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.

[16] A. Queiroz e R. Coelho, "Noisy Speech Based Temporal Decomposition to Improve Fundamental Frequency Estimation," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, v. 30, pp. 2504–2513, 2022.

[17] 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.

[18] M. Jeub, M. Schaefer, e P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *2009 16th International Conf. on Digital Signal Processing.*, pp. 1-5, Jul. 2009.

[19] 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.

[20] R. Coelho e 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, e C.Cavalcante: Eds. Boca Raton, FL, USA: CRC Press, 2015.

[21] S. Gonzalez, e M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," *Proc. 19th Eur. Signal Process. Conf.*, pp. 451-455, 2011.

[22] I. R. Titze, "*Principles of Voice Production*,"Eiglenwood Cliffs, NJ, USA: Prentice Hall, 1994.

[23] D. Talkin, W. R. Klein, e K. Paliwal, "A robust algorithm for pitch tracking," *Speech Coding and Synthesis*, Amsterdam, Netherlands: Elsevier, 1995, pp. 497-518.

[24] 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.

[25] 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.

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