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

Anderson Queiroz da Silva
Rosângela Fernandes Coelho

Laboratório de Processamento de Sinais Acústicos
Instituto Militar de Engenharia

XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT

26-29 de Setembro de 2021

Plano de Apresentação

- Motivação e Desafios
- Objetivos
- Frequência Fundamental - Estimação e Métricas de avaliação
- Aprendizado DCNN (*Deep Convolutional Neural Network*) para classificação dos quadros em Alta/Baixa Frequência.
  - Contribuição para a correção dos erros de estimação da F_0
- Experimentos e Resultados
  - Cenário Experimental
  - Resultados de Erro GE (*Gross Error*) e MAE (*Mean Absolute Error*)
- Conclusão

Motivação

Motivação

- Ruídos Acústicos em cenários urbanos:
  1. Afetam a comunicação pela fala entre seres humanos
  2. Provoca queda na eficiência de sistemas de reconhecimento da fala
  3. Compromete o aprendizado em salas de aula
  4. Mascaram os componentes e estruturas da voz, como a Frequência Fundamental

Motivação

- Aplicações da Estimação da F_0:
  1. Sistemas de Identificação de Locutor
  2. Síntese e Análise de sinais de voz
  3. Música
  4. Detecção de Distúrbios da voz

Motivação

- Aplicações da Estimação da F_0:
  1. Sistemas de Identificação de Locutor
  2. Síntese e Análise de sinais de voz
  3. Música
  4. Detecção de Distúrbios da voz
- Desafios: Diversidade de ambientes, não-estacionaridade dos sinais acústicos, mascaramento dos componentes harmônicos da voz.

Objetivos

- Propor uma composição de aprendizado de máquina DCNN + HHT-Amp para aprimorar a acurácia das estimativas de frequência fundamental.

Objetivos

- Propor uma composição de aprendizado de máquina DCNN + HHT-Amp para aprimorar a acurácia das estimativas de frequência fundamental.
- Avaliar a composição DCNN com outros métodos de estimação da F_0 presentes na literatura.

Objetivos

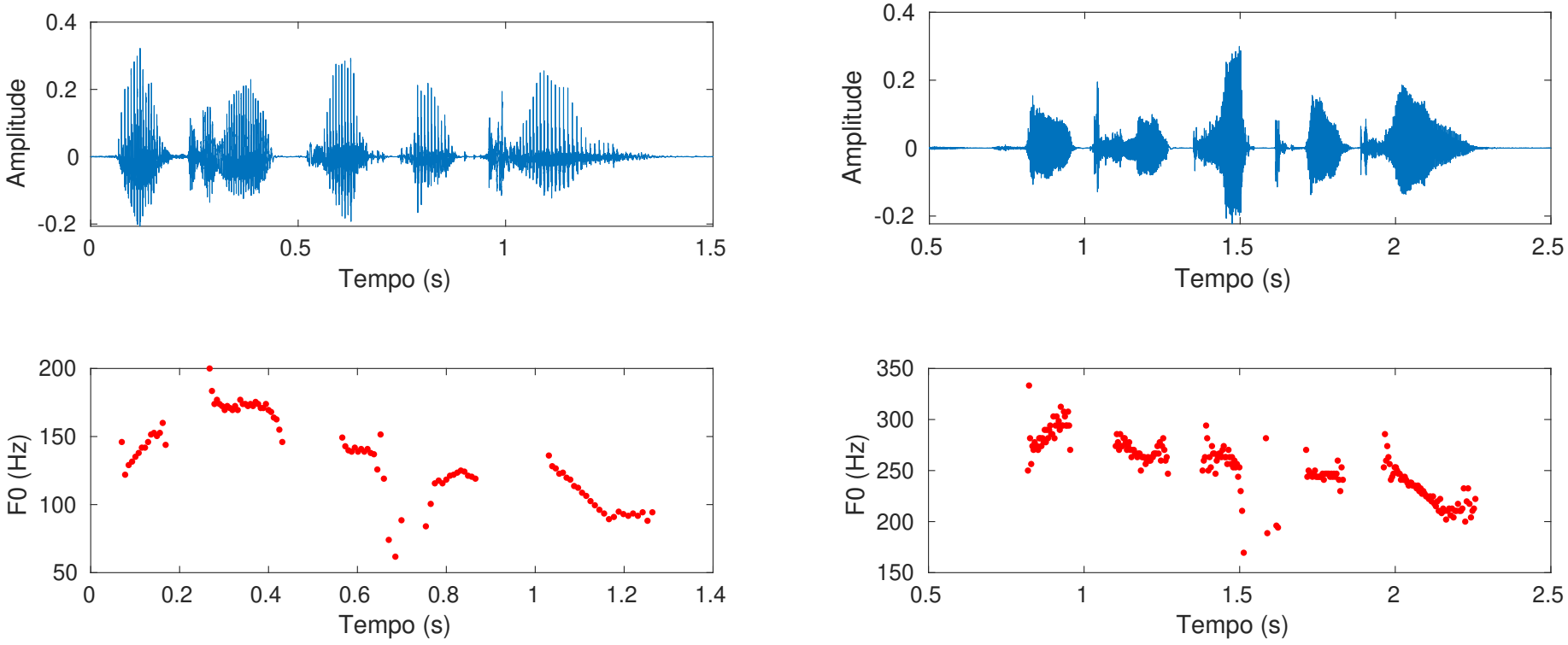
- Propor uma composição de aprendizado de máquina DCNN + HHT-Amp para aprimorar a acurácia das estimativas de frequência fundamental.
- Avaliar a composição DCNN com outros métodos de estimação da F_0 presentes na literatura.
- Avaliar a capacidade de aprimoramento da acurácia da F_0 em ambientes ruidosos.

Frequência Fundamental de Sinais de Voz

- Relacionada com as características biométricas inerentes ao trato vocal de seres humanos.

Frequência Fundamental de Sinais de Voz

- Relacionada com as características biométricas inerentes ao trato vocal de seres humanos.



Estimadores Presentes na Literatura

- ACF (*Autocorrelation Function*)*:
  - Estimador clássico, baseado na Função Autocorrelação

Estimadores Presentes na Literatura

- ACF (*Autocorrelation Function*)*:
  - Estimador clássico, baseado na Função Autocorrelação
- SHR (*Subharmonic-to-Harmonic Ratio*)**:
  - Análise das características espectrais do sinal de voz

Estimadores Presentes na Literatura

- ACF (*Autocorrelation Function*)*:
  - Estimador clássico, baseado na Função Autocorrelação
- SHR (*Subharmonic-to-Harmonic Ratio*)**:
  - Análise das características espectrais do sinal de voz
- SFF (*Single Frequency Filtering*)***:
  - Autocorrelação dos envelopes com maior concentração de energia.

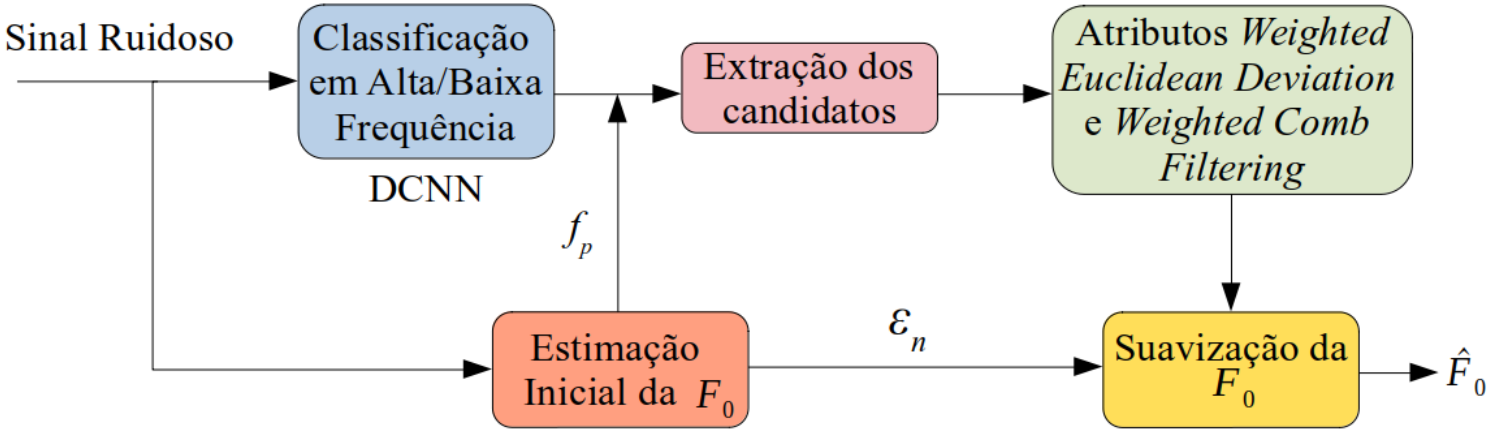
Estimadores Presentes na Literatura

- ACF (*Autocorrelation Function*)*:
  - Estimador clássico, baseado na Função Autocorrelação
- SHR (*Subharmonic-to-Harmonic Ratio*)**:
  - Análise das características espectrais do sinal de voz
- SFF (*Single Frequency Filtering*)***:
  - Autocorrelação dos envelopes com maior concentração de energia.
- HHT-Amp (*Hilbert-Huang Transform*)****
  - Atuação no domínio do tempo em sinais ruidosos.

* [L. Rabiner, "On the use of autocorrelation analysis for pitch detection," IEEE Trans. Acoust., Speech Signal Process., v. 25, pp. 24-33, Feb. 1977.]
** [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.]
*** [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.]
**** [L. Zão e R. Coelho, "On the estimation of fundamental frequency from nonstationary noisy speech signals based on the Hilbert-Huang Transform," IEEE Signal Process. Lett., v. 25, no. 2 pp. 248-252, Feb. 2018.]

Aprendizado DCNN e sua importância na correção dos erros nas Estimativas da F_0

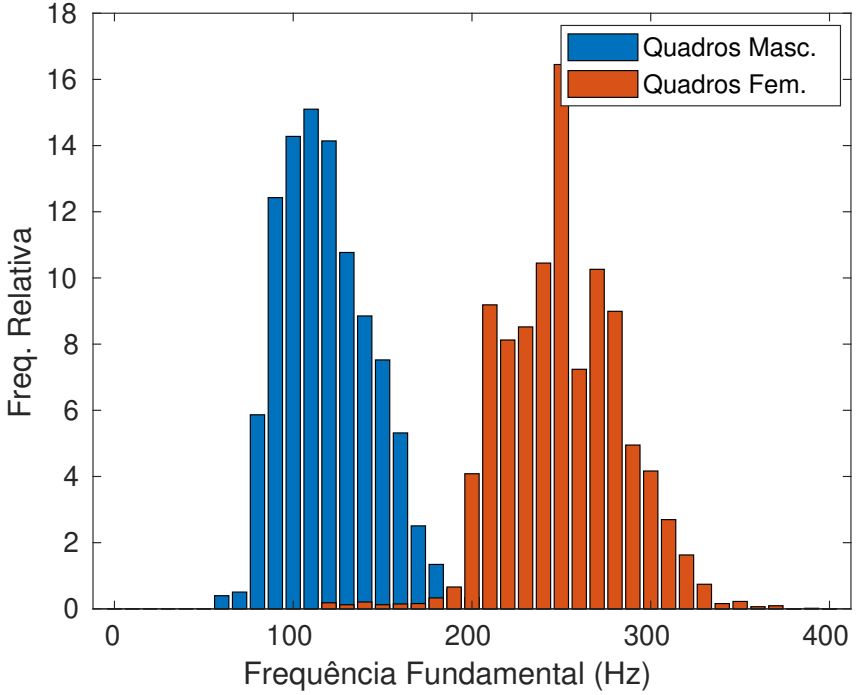
- Diagrama Esquemático da Proposta*



* [Adaptado de: 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.]

Definição do limiar de Alta/Baixa Frequência.

- Frequência Relativa da F_0 dos sinais de voz da base CSTR*.
- Limiar de Baixa/Alta Frequência definido em 200 Hz.



* [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. EUROSPEECH93., pp. 1003-1006, Sep. 1993.]

Extração dos Candidatos

- Tabela com os Candidatos extraídos a partir da classificação em Alta/Baixa Frequência.

	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	$\{0{,}5f_p\}$	-

Definição da F_0 Aprimorada

- Extração de Atributos Espectrais:
  - WED (*Weighted Euclidean Deviation*) - $d_{j,n}$
  - WCF (*Weighted Comb Filtering*) - $y_{j,n}$

Definição da F_0 Aprimorada

- Extração de Atributos Espectrais:
  - WED (*Weighted Euclidean Deviation*) - $d_{j,n}$
  - WCF (*Weighted Comb Filtering*) - $y_{j,n}$

- Função Custo para cada Candidato:

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

Definição da F_0 Aprimorada

- Extração de Atributos Espectrais:
  - WED (*Weighted Euclidean Deviation*) - $d_{j,n}$
  - WCF (*Weighted Comb Filtering*) - $y_{j,n}$

- Função Custo para cada Candidato:

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

onde:

$|\log f_{j,n} - \log f_{i,n+1}|$: distância entre candidatos de quadros vizinhos;

λ e α são parâmetros de regularização, de 1,4 e 1,7, respectivamente.

pr_n: probabilidade da camada de saída (*Softmax*) da DCNN.

$\varepsilon_n = 1$ se $f_{j,n} = f_p$ e igual a zero nos demais casos.

Medidas de Erro de Estimação da F_0

- GE - *Gross Error Rate**
  - GE $= (E_{F0}/N_{F0}) * 100$, onde N_{F0} é a quantidade total de estimações e E_{F0} a parcela que satifazem a condição $|(\hat{F}_0/F_0) - 1| > 0,2$

Medidas de Erro de Estimação da F_0

- GE - *Gross Error Rate**
  - GE $= (E_{F0}/N_{F0}) * 100$, onde N_{F0} é a quantidade total de estimações e E_{F0} a parcela que satifazem a condição $|(\hat{F}_0/F_0) - 1| > 0,2$

- MAE - *Mean Absolute Error**
  - MAE $= \left(\sum_{i=1}^{n} \left|\hat{F}_0(i) - F_0(i)\right|\right)/n$, onde $\hat{F}_0(i)$ é a estimação e $F_0(i)$ a referência.

* [L. Rabiner, M. Cheng, A. Rosenberg and C. McGonegal, "A comparative performance study of several pitch detection algorithms," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 5, pp. 399-418, October 1976.]

Cenário Experimental

- Bases de Sinais de Voz Adotadas:
  - 100 locuções da base de voz CSTR
  - 192 locuções da base de voz TIMIT*
- Ruídos Adotados:
  - Balbúrdia e Volvo (base RSG-10**); Trem, Cafeteria e Helicóptero (Freesound.org); e SSN (*Speech Shaped Noise*) da base DEMAND***
  - Valores de SNR: -10dB, -5dB, 0dB e 5dB.
- Processo de Treinamento
  - 30% da base CSTR e 30% da TIMIT.
  - Total de 197.130 quadros de 960 amostras.
  - SNRs -10dB e 0dB.

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

Ruídos Acústicos - Análise

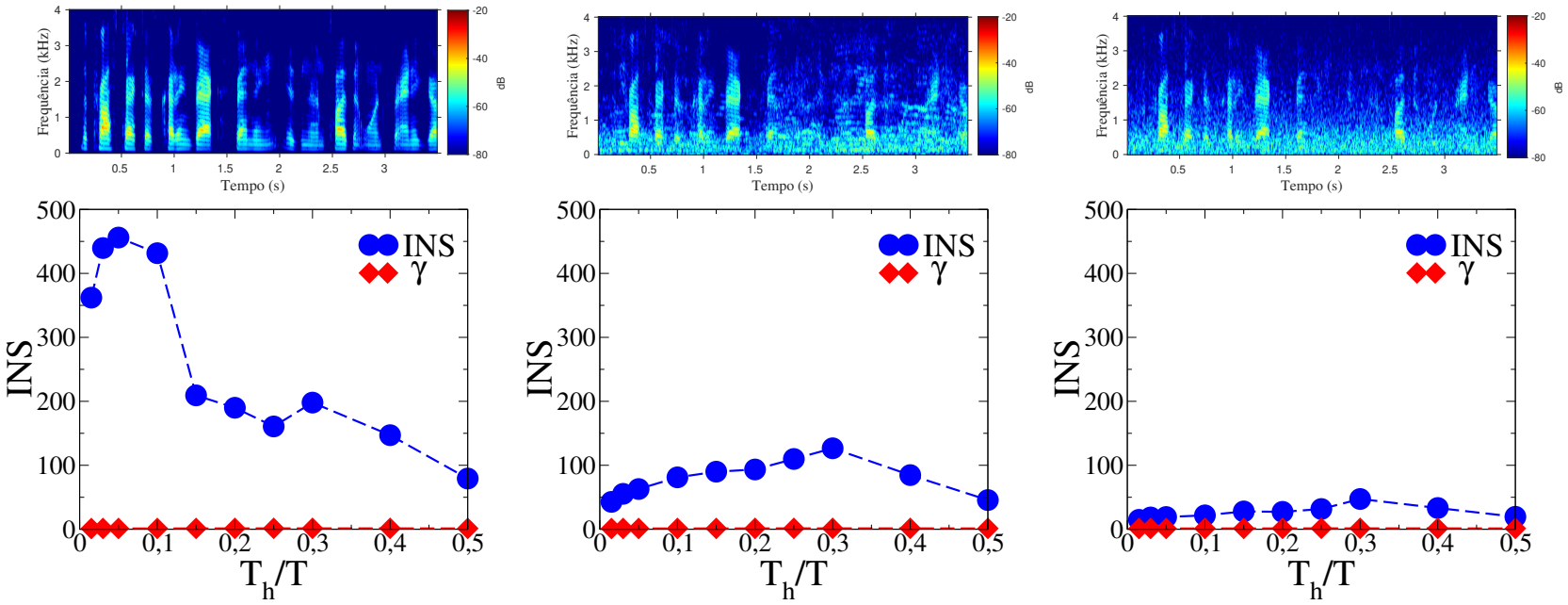
- Análise da Não-estacionaridade: INS* - *Index of Non-Stationarity*.
- Medida objetiva baseada na análise tempo-frequência
- Divergência de Kullback-Leibler (KL) mede distância entre espectro de tempo-curto (T_h) e espectro global (T)
- INS é a razão entre KLs do sinal e o valor correspondente aos referenciais estacionários surrogates

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

* [Testing Stationarity With Surrogates: A Time-Frequency Approach, P. Borgnat, P. Flandrin, P. Honeine, C. Richard and J. Xiao, IEEE Transactions on Signal Processing, vol. 58, no. 7, Jul 2010.]

INS dos Sinais de Voz Ruidosos

- INS do sinal de voz Limpo e Ruidosos com: Balbúrdia e SSN com SNR = 0dB.



$INS_{máx} = 456$ $INS_{máx} = 127$ $INS_{máx} = 47$

Resultados Iniciais de GE e MAE

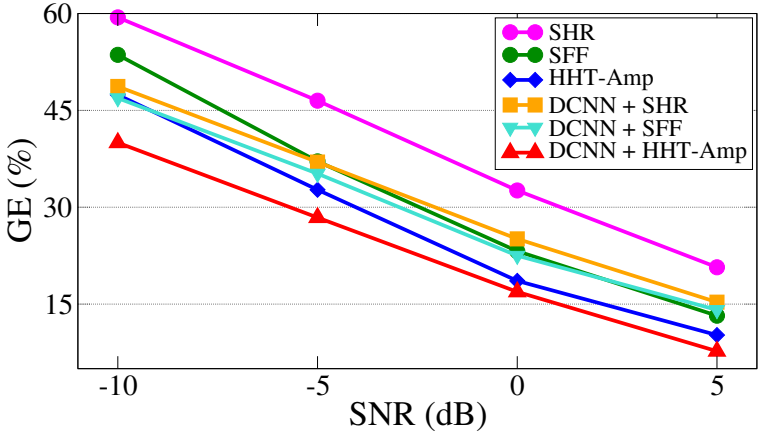
Ruído	SNR	GE (%)			MAE (Hz)		
		SHR	SFF	HHT-Amp	SHR	SFF	HHT-Amp
Balbúrdia $INS_{máx}$=34,6	-10 dB	59,4	53,6	**47,4**	53,9	45,1	**41,4**
	-5 dB	46,5	37,1	**32,7**	43,7	32,6	**29,3**
	0 dB	32,6	23,2	**18,6**	32,5	22,9	**18,1**
	5 dB	20,7	13,2	**10,2**	22,7	16,0	**11,0**
	Média	39,8	31,8	**27,2**	38,2	29,2	**24,9**
Trem $INS_{máx}$=18,8	-10 dB	61,2	35,6	**30,3**	65,4	32,8	**29,3**
	-5 dB	46,4	21,0	**18,7**	50,4	21,5	**18,1**
	0 dB	32,4	11,0	**9,3**	36,5	14,1	**10,2**
	5 dB	21,1	6,4	**5,2**	25,2	10,9	**6,8**
	Média	40,3	18,5	**15,9**	44,4	19,8	**16,1**
Cafeteria $INS_{máx}$=11,7	-10 dB	61,6	47,4	**44,9**	69,2	45,1	**40,2**
	-5 dB	48,6	32,4	**29,5**	55,9	32,0	**27,7**
	0 dB	34,0	20,7	**17,0**	40,4	22,1	**17,2**
	5 dB	21,6	13,1	**9,0**	27,2	16,2	**10,4**
	Média	41,4	28,4	**25,1**	48,2	28,9	**23,9**

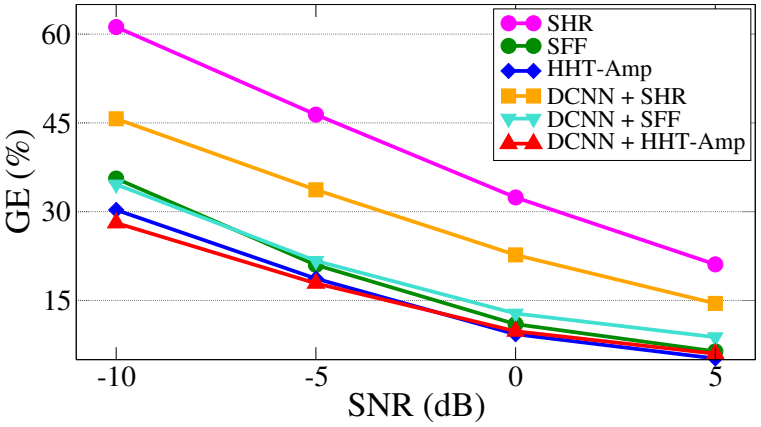
Resultados Iniciais de GE e MAE

Ruído	SNR	GE (%)			MAE (Hz)		
		SHR	SFF	HHT-Amp	SHR	SFF	HHT-Amp
Helicóptero $INS_{máx}$=1,8	-10 dB	73,7	48,1	**33,1**	74,6	44,6	**31,1**
	-5 dB	56,4	26,2	**18,1**	58,0	26,5	**18,1**
	0 dB	38,4	15,1	**9,9**	40,7	17,4	**10,7**
	5 dB	24,4	8,3	**5,4**	27,3	12,2	**7,1**
	Média	48,2	24,4	**16,6**	50,2	25,1	**16,7**
SSN $INS_{máx}$=1,6	-10 dB	68,0	61,1	**49,8**	64,5	58,0	**43,9**
	-5 dB	53,7	39,8	**33,7**	52,6	39,3	**30,6**
	0 dB	37,0	21,0	**19,1**	38,3	23,2	**19,1**
	5 dB	23,7	11,3	**9,5**	26,2	15,4	**10,5**
	Média	45,6	33,3	**28,0**	45,4	34,0	**26,0**
Volvo $INS_{máx}$=0,9	-10 dB	30,9	19,6	**7,2**	56,0	21,1	**10,3**
	-5 dB	20,9	12,7	**4,4**	38,3	15,6	**6,8**
	0 dB	13,6	8,5	**3,1**	25,3	12,6	**5,3**
	5 dB	9,2	5,8	**2,9**	17,0	10,1	**4,8**
	Média	18,7	11,6	**4,4**	34,2	14,8	**6,8**
Média Total		38,1	23,4	**18,3**	43,0	24,4	**17,5**

Resultados de GE

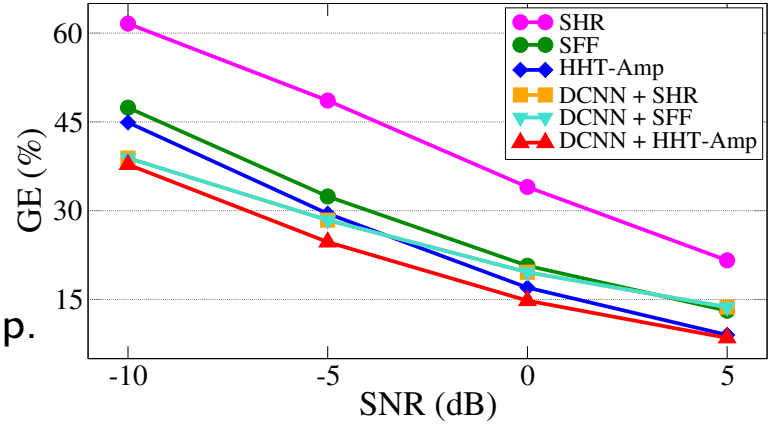
- Sinais corrompidos pelos ruídos Balbúrdia, Trem e Cafeteria.





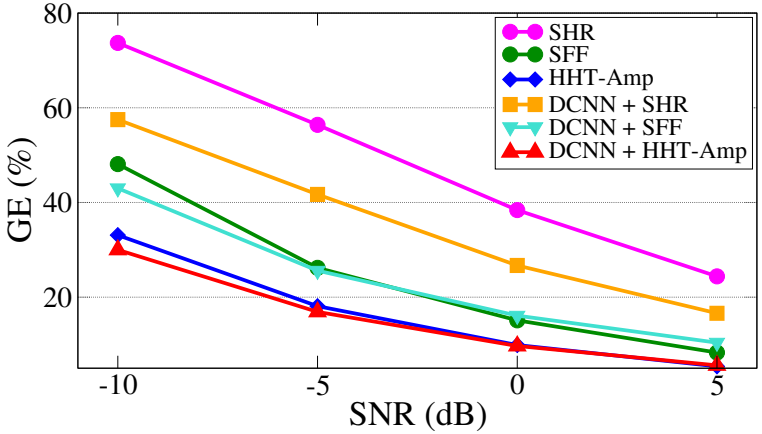
- Balbúrdia (-10 dB)

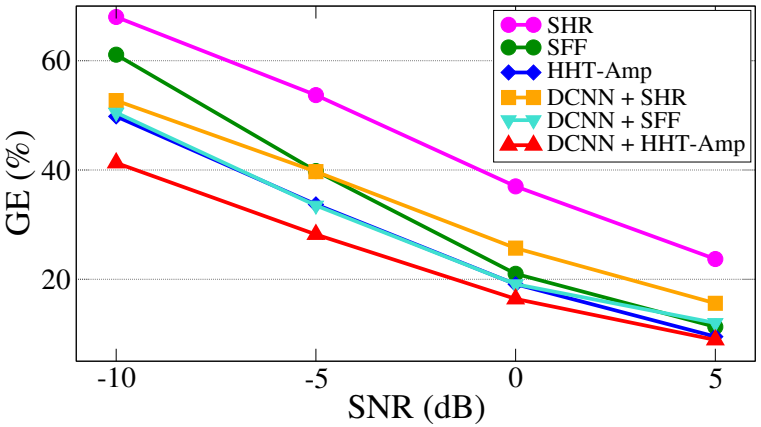
DCNN+SHR $\Rightarrow \downarrow$ 10,7p.p.
DCNN+SFF $\Rightarrow \downarrow$ 6,8p.p.
DCNN+HHT-Amp $\Rightarrow \downarrow$ 7,4p.p.



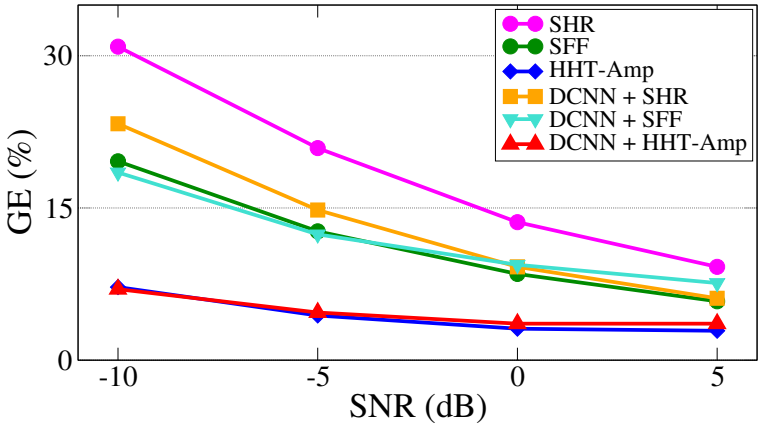
Resultados de GE

- Sinais corrompidos por Helicóptero, SSN e Carro.





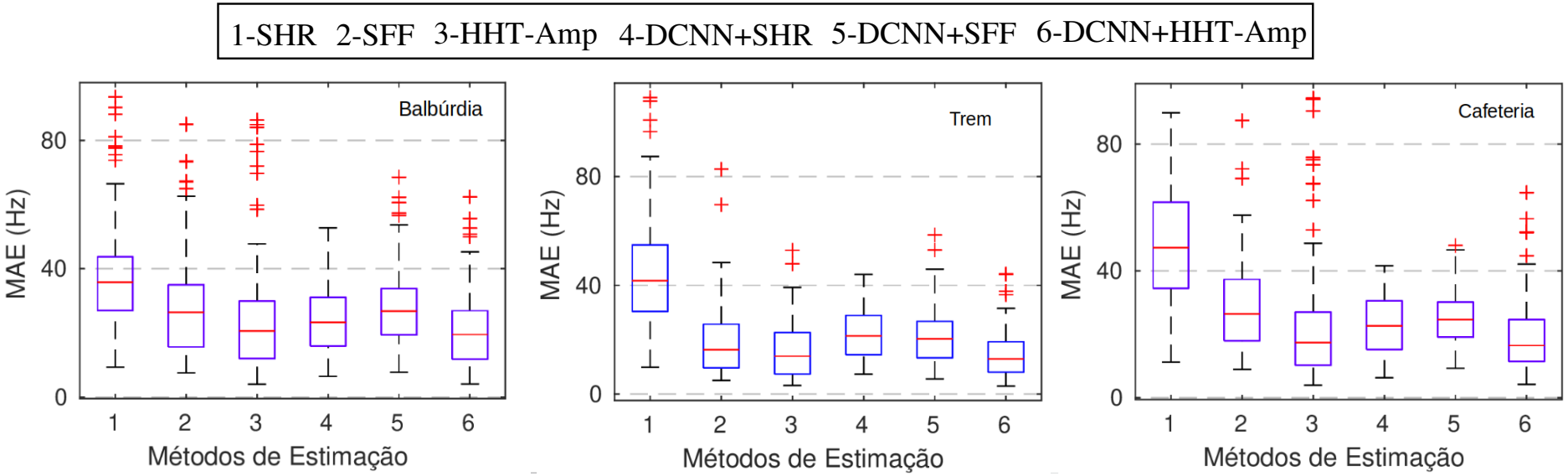
- DCNN+HHT-Amp
  SNR = -10 dB



Balbúrdia ⇒ GE = 40,0%.
Helicóptero ⇒ GE = 30,0%.
Carro ⇒ GE = 7,0%.

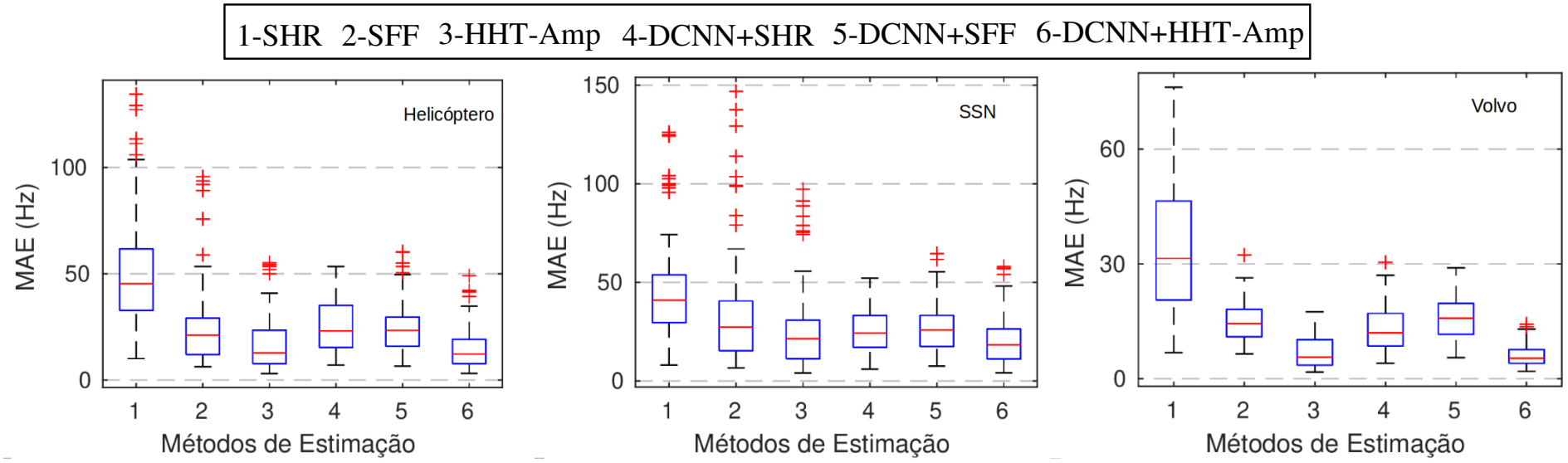
Resultados de MAE

- Sinais de voz ruidosos com Balbúrdia, Trem e Cafeteria.
- Distribuição dos resultados para os quatro valores de SNR (-10 dB, -5 dB, 0 dB e 5 dB).



Resultados de MAE

- Sinais de voz corrompidos por Helicóptero, SSN e Volvo.



Conclusão

- O Aprendizado DCNN aplicado na classificação dos quadros em Alta/Baixa Frequência, contribuiu na detecção e correção de erros de estimação da Frequência Fundamental.

- A composição DCNN + HHT-Amp atingiu melhor acurácia dentre os métodos comparativos, inclusive nas condições mais severas de ruídos, reduzindo as taxas de GE em até 17%.

Agradecimentos

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

Conselho Nacional de Desenvolvimento Científico e Tecnológico

Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro

OBRIGADO!