XXXIII Simpósio Brasileiro de Telecomunicações
1 a 4 de Setembro de 2015 - Juiz de Fora - MG - Brasil
Sociedade Brasileira de Telecomunicações
Universidade Federal de Juiz de Fora

# Análise da Não-Estacionariedade de Sinais Acústicos e seu Efeito na Predição Objetiva da Inteligibilidade: Desafios e Soluções

**Rosângela Fernandes Coelho e Guilherme Zucatelli Nossa**
**Laboratório de Processamento de Sinais Acústicos**
**Instituto Militar de Engenharia (IME), Rio de Janeiro, Brasil**
**coelho@ime.eb.br**

## Resumo

Sistemas baseados na análise e classificação de sinais acústicos têm ampla aceitação em diferentes áreas e aplicações como: autenticação biométrica de transações bancárias, biomedicina, terapia com análise de emoções acústicas, análise de patologias vocais, monitoramento de sinais sísmicos, localização de fontes acústicas, reconhecimento de música em ambiente e identificação de indivíduos.

## Interferência e Variações Acústicas

Um grande desafio para a pesquisa é ocasionado quando os sinais de voz a serem processados são capturados em ambientes com presença de interferência de outras fontes acústicas (avião, trem, carro, arma de fogo) ou por sinais que apresentam alterações acústicas decorrentes de estados emocionais (raiva, medo, tristeza, felicidade). Neste cenário, além da qualidade, a inteligibilidade sonora, definida como medida que reflete o quanto uma mensagem acústica é compreensível, pode ser muito afetada. E, consequentemente, acarretar em severa degradação na acurácia dos sistemas de classificação acústica pelo sinal de voz.

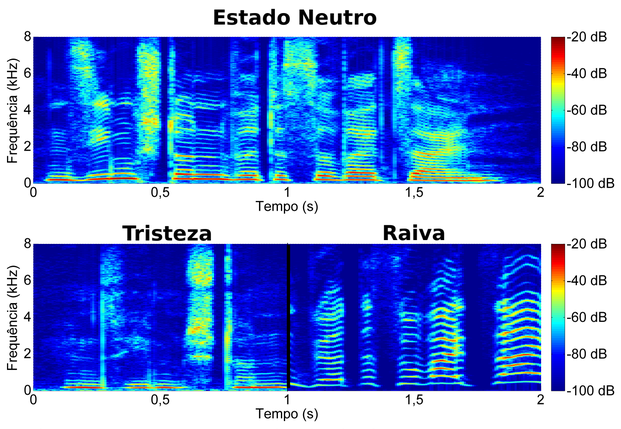


Fig. 1.: Espectrogramas de sinais de voz em estado neutro e com variações acústicas emocionais.

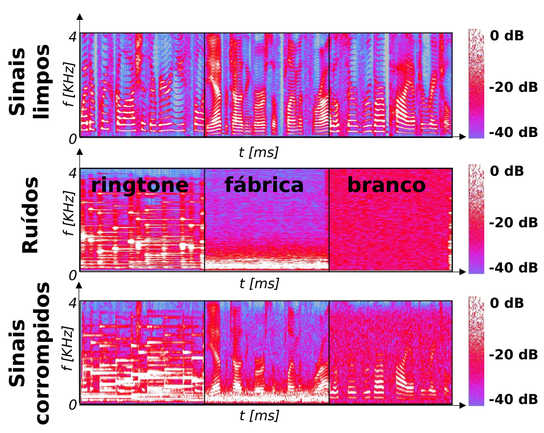


Fig. 2: Espectrogramas de sinais de voz limpos, ruídos, e os mesmos sinais de voz corrompidos.

**Abordagem:**
- Quantificar o Índice de Não-Estacionariedade (INS).
- Medir a qualidade e a inteligibilidade sonora.

## Índice de Não-Estacionariedade

O índice de não-estacionariedade (INS - *index of non-stationarity*) é uma medida tempo-frequência que reflete o grau de não-estacionariedade de um sinal. Esta avaliação é realizada a partir da comparação das componentes espectrais do sinal e de referenciais (*surrogates*) estacionários. Estas componentes são obtidas aplicando-se a transformada de Fourier de tempo curto para um determinado tamanho da janela ($T_h$). A distância de Kullback-Leibler (KL) é utilizada para medir a oscilação das componentes espectrais do sinal ao longo do tempo. Finalmente, o INS é definido como a razão entre os valores de KL obtidos do sinal analisado e doss seus referenciais estacionários. Após a representação dos valores de KL por uma distribuição Gama, é definido um limiar ($\gamma$) para o teste:

$$\mathrm{INS} \begin{cases} \leqslant \gamma, & \text{inal e tacionário} \\ > \gamma, & \text{inal não-e tacionário} \end{cases}$$

Os valores de $\gamma$ são definidos considerando um grau de confiança de 95%.

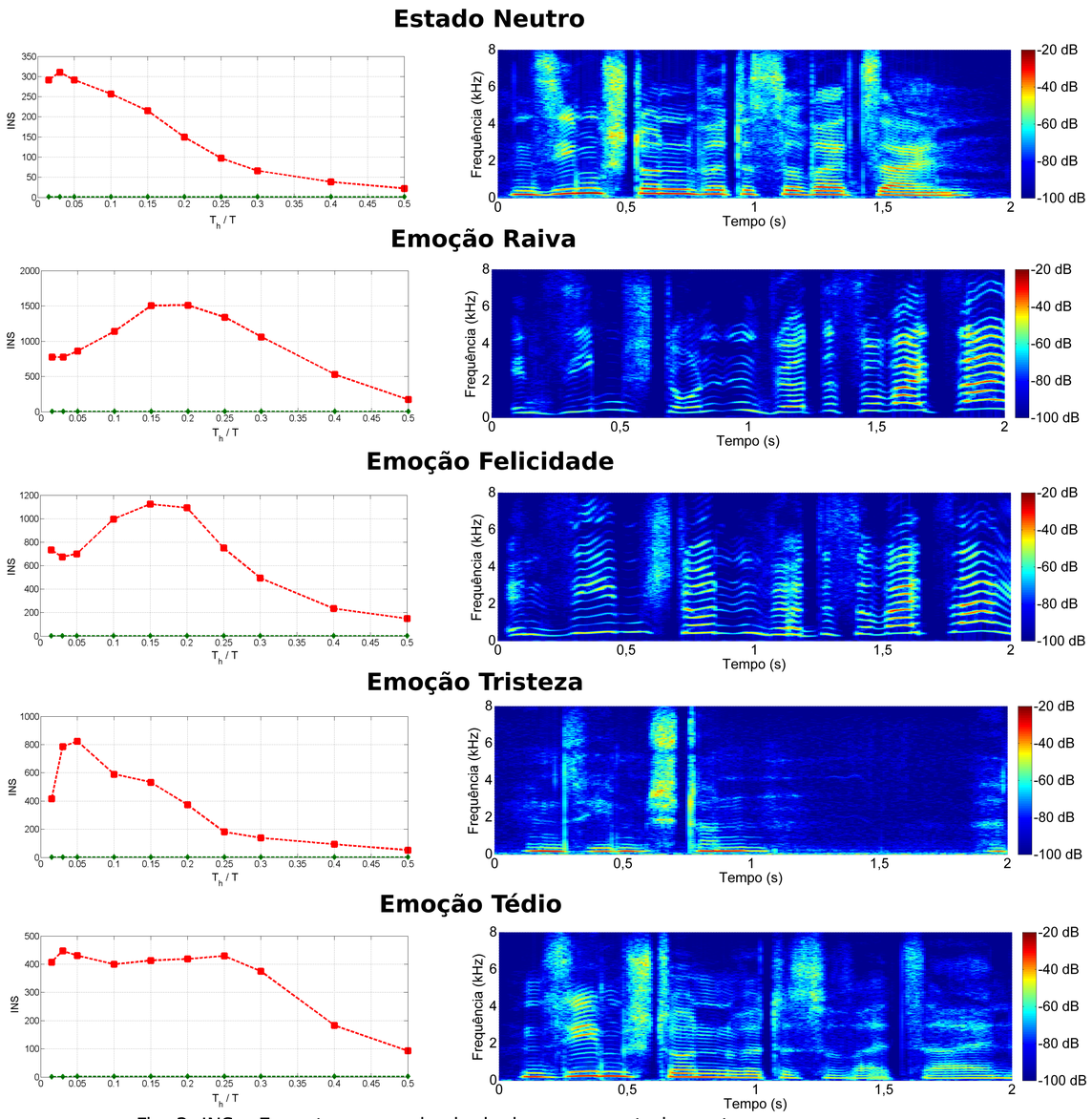


Fig. 3: INS e Espectrograma de sinais de voz em estado neutro e com presença de múltiplas variações emocionais.

## Inteligibilidade

Métodos de realce que apresentam bons resultados na predição da qualidade dos sinais realçados, geralmente definida como medida do desconforto causado por um ruído acústico ao ouvinte, não necessariamente atingem uma boa inteligibilidade sonora. O resultado desta predição objetiva deve ser fortemente correlacionado com os obtidos de testes perceptuais. Este aspecto influencia a classificação de sinais de voz e, consequentemente, o desempenho das aplicações relacionadas.

**Medidas:**
- **CSII** (*Coherence Speech Intelligibility Index*)
- **fAI** (*Fractional Articulation Index*)
- **STOI** (*Short-Time Objective Intelligibility measure*)

## Soluções

As soluções devem processar e tratar interferências acústicas (ruídos e emoções) para atenuar seus efeitos. Como?

- **Ruídos → Técnicas de Realce:**
  - **Espectrais**
    - **SS** (*Spectral Subtraction*)
    - **UMMSE/Wiener** (*Unbiased minimum mean-square error*)
    - **IMCRA/OMLSA** (*Improved Minima-Controlled Recursive Averaging / Optimally Modified Log-Spectral Amplitude*)
  - **Temporais**
    - **EMD-DT** (*Empirical Mode Decomposition Detrending*)
    - **EMDF** (*EMD-based Filtering Post-Enhancement*)
    - **EMDH** (*EMD with Hurst-based Mode Selection*)
- **Emoções → Detecção de Múltiplas Emoções**

## Bases: Voz, Ruídos e Emoções

**Voz:**
- KING
- TIMIT
- NTIMIT
- YOHO

**Ruídos:**
- NOISEX-92
- AURORA
- SPINE
- MIT

**Emoções :**
- EMO-DB (Alemão)
- SUSAS (Ingês)
- GEMEP (Francês)

## Resultados

- Qualidade e Inteligibilidade

| Ruído | SNR | SS | OMLSA | Wiener | EMDH |
|---|---|---|---|---|---|
| Balbúrdia | 10 | 89.6 | 89.5 | 88.8 | 89.0 |
| | 5 | 69.8 | 72.7 | 72.0 | 73.4 |
| | 0 | 28.8 | 36.6 | 37.6 | 42.0 |
| | -5 | 5.9 | 8.1 | 9.2 | 12.5 |
| | -10 | 1.1 | 1.1 | 1.6 | 2.6 |
| | Média | 39.0 | 41.6 | 41.8 | 43.9 |
| Britadeira | 10 | 90.7 | 91.9 | 92.9 | 92.7 |
| | 5 | 71.0 | 81.2 | 85.9 | 86.9 |
| | 0 | 37.9 | 61.5 | 72.2 | 72.4 |
| | -5 | 13.9 | 26.8 | 44.1 | 42.4 |
| | -10 | 5.2 | 7.0 | 17.4 | 16.5 |
| | Média | 43.7 | 53.7 | 62.5 | 62.2 |
| Serra Elétrica | 10 | 86.6 | 85.6 | 85.7 | 88.2 |
| | 5 | 55.1 | 56.4 | 57.0 | 61.8 |
| | 0 | 16.7 | 16.0 | 19.3 | 25.1 |
| | -5 | 2.8 | 2.3 | 3.5 | 5.0 |
| | -10 | 0.7 | 0.7 | 1.1 | 1.5 |
| | Média | 32.4 | 32.2 | 33.3 | 36.3 |
| Trem | 10 | 90.8 | 90.5 | 90.2 | 90.0 |
| | 5 | 81.6 | 81.0 | 80.8 | 81.0 |
| | 0 | 60.1 | 63.1 | 63.5 | 63.6 |
| | -5 | 23.4 | 34.9 | 35.9 | 35.3 |
| | -10 | 4.8 | 9.4 | 11.2 | 10.8 |
| | Média | 52.1 | 55.8 | 56.3 | 56.1 |

Tab. 1: Resultados de inteligibilidade com STOI.

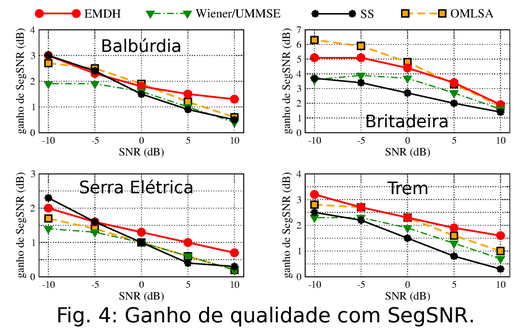


Fig. 4: Ganho de qualidade com SegSNR.

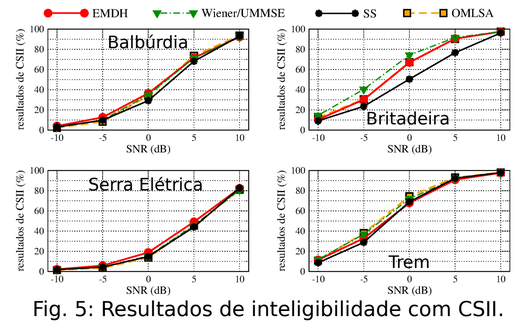


Fig. 5: Resultados de inteligibilidade com CSII.

- Detecção de Múltiplas Emoções

| | Desgosto | Felicidade | Medo | Raiva | Neutro | Tédio | Tristeza |
|---|---|---|---|---|---|---|---|
| Desgosto | **67** | 10 | 6 | 0 | 10 | 7 | 0 |
| Felicidade | 6 | **48** | 8 | 25 | 11 | 2 | 0 |
| Medo | 5 | 16 | **62** | 0 | 11 | 3 | 3 |
| Raiva | 2 | 10 | 2 | **86** | 0 | 0 | 0 |
| Neutro | 2 | 0 | 8 | 0 | **71** | 17 | 2 |
| Tédio | 13 | 0 | 2 | 0 | 20 | **61** | 4 |
| Tristeza | 0 | 0 | 0 | 0 | 6 | 12 | **82** |
| **Média de acertos: 68,1 %** | | | | | | | |

Tab. 2: Matriz de confusão com o atributo pH.

| | Desgosto | Felicidade | Medo | Raiva | Neutro | Tédio | Tristeza |
|---|---|---|---|---|---|---|---|
| Desgosto | **61** | 11 | 5 | 5 | 18 | 0 | 0 |
| Felicidade | 8 | **58** | 11 | 19 | 3 | 2 | 0 |
| Medo | 11 | 22 | **33** | 7 | 15 | 13 | 0 |
| Raiva | 1 | 14 | 0 | **85** | 0 | 0 | 0 |
| Neutro | 5 | 0 | 5 | 0 | **65** | 23 | 1 |
| Tédio | 8 | 5 | 5 | 0 | 25 | **53** | 4 |
| Tristeza | 11 | 0 | 0 | 0 | 9 | 6 | **74** |
| **Média de acertos: 61,3 %** | | | | | | | |

Tab. 3: Matriz de confusão com o atributo MFCC.

## Desafios e Tendências

- Definir métodos de realce que considerem não somente a qualidade mas a inteligibilidade dos sinais.
- Definir medidas de predição de inteligibilidade objetiva sem conhecimento prévio de um sinal limpo.
- Inexistência de soluções para detecção robusta de múltiplas emoções.

## Referências Destaque


- T. Quatieri, *Discrete-Time Speech Signal Processing.* Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
- P. Borgnat, P. Flandrin, P. Honeine, C. Richard e J. Xiao, "Testing Stationarity With Surrogates: A Time-Frequency Approach," IEEE Transactions on Signal Processing, vol.58, no.7, pp. 3459-3470, Julho 2010.
- P. Loizou, *Speech Enhancement: Theory and Practice. Boca Raton, FL, USA: CRC Press, 2013.*
- P. Loizou e K. Gibak, "Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 1, pp. 47-56, Janeiro 2011.
- C. Taal, R. Hendriks, R. Heusdens e J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4214-4217, 14-19 Março 2010.
- T. Gerkmann e 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, no. 4, pp. 1383-1393, Maio 2012.
- L. Zão, R. Coelho e P. Flandrin, "Speech Enhancement with EMD and Hurst-Based Mode Selection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 5, pp. 899-911, Maio 2014.
- B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll e A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," IEEE Workshop on Automatic Speech Recognition & Understanding, pp.552-557, 2009.
- L. Zão, D. Cavalcante, R. Coelho, "Time-Frequency Feature and AMS-GMM Mask for Acoustic Emotion Classification," IEEE Signal Processing Letters, vol. 21, no. 5, pp. 620-624, Maio 2014.
- R. Sant Ana, R. Coelho and A. Alcaim, "Text-Independent Speaker Recognition Based on the Hurst Parameter and the Multi-Dimensional Fractional Brownian Motion Model", IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 3, pp. 931-940, Maio 2006.