# Estudo da Inteligibilidade de Sinais com Interferência Acústica de *Drones*

V. Freitas, R. Coelho, P. Rosa

***Resumo*— Este trabalho investiga o efeito de ego-interferências sobre a inteligibilidade de sinais de voz captados por *drones*. Três máscaras acústicas são utilizadas para prover ganho de inteligibilidade aos sinais de voz considerando dois cenários com reverberação e um cenário sem reverberação. A análise inclui ainda a comparação entre os impactos das ego-interferências e de ruídos ambientais reais com razão sinal-ruído variando de -10 dB a 5 dB. Duas medidas de predição de inteligibilidade de voz são adotadas para avaliar objetivamente as máscaras acústicas. Os resultados mostram que a máscara binária IRM, desenvolvida para sinais reverberantes, proporciona maior aprimoramento médio que a IBM em todos os cenários. Além disso, a máscara suave IRatioM alcança inteligibilidade superior em relação às máscaras binárias.**

***Palavras-Chave*— Inteligibilidade, *Drones*, Ego-interferência, Reverberação, Máscaras acústicas, Medidas objetivas.**

***Abstract*— This work investigates the ego-noise effects on the intelligibility of speech signals captured by drones. Three acoustic masks are applied to improve speech intelligibility considering two reverberant scenarios and a non-reverberant scenario as well. The analysis also includes the comparison of the impacts of ego-noises and ambient noises with signal-to-noise ratio varying from -10 dB to 5 dB. Two speech intelligibility prediction measures are adopted to objectively evaluate the acoustic masks. Results show that the binary IRM mask, which was specially developed for reverberant signals, outperforms the IBM in terms of intelligibility gain for all scenarios. Furthermore, the soft mask IRatioM achieves the best intelligibility results.**

***Keywords*— Speech intelligibility, Drones, Ego-noise, Reverberation, Acoustic masks, Objective measures.**

## I. Introdução

O sinal de voz é comumente capturado em ambientes urbanos acusticamente ruidosos: um cenário desafiador na presença de *drones* com ego-interferências (EN, *ego-noises*) [1]. Ilustrações desse cenário são apresentadas na Figura 1. Ego-interferências são variações acústicas próprias, causadas pelos motores rotativos e pelas hélices dos *drones* [2]. Sinais captados por sensores embarcados em *drones* são altamente degradados por EN. Nestes casos, a razão sinal-ruído (SNR, *signal-to-noise ratio*) é frequentemente inferior a -10 dB [3].

Nos últimos anos, o uso de *drones* foi intensificado em atividades de busca e salvamento, monitoramento e interação humano-*drone* [4][5]. Em particular, técnicas de *drone audition* visam localizar fontes sonoras e interpretar sinais acústicos no ambiente ao redor do *drone*. Sistemas de reconhecimento de voz, por exemplo, podem ser altamente impactados pela presença de interferências acústicas que afetam a inteligibilidade [6]. Portanto, a análise da inteligibilidade acústica de sinais captados em ambientes reverberantes e ruidosos é essencial para solucionar desafios *drone audition*.

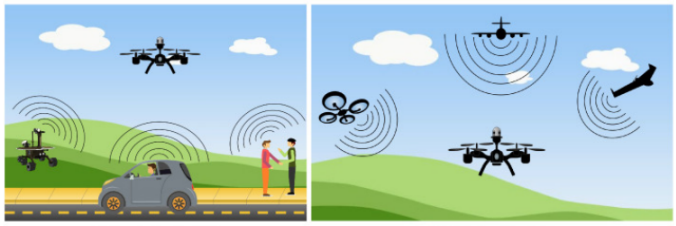

Fig. 1. Sensores acústicos embarcados em *drones* captam sinais de fontes localizadas no solo ou no ar. [2]

Máscaras acústicas [7][8][9] foram propostas na literatura para aprimoramento da inteligibilidade de voz em ambientes urbanos. As máscaras emulam a capacidade humana de direcionar atenção para uma única fonte, mesmo na presença de um ou mais sinais concorrentes. Este efeito é conhecido como *Cocktail Party*. Máscaras ideais, como a IBM (*Ideal Binary Mask*) [10], a IRatioM (*Ideal Ratio Mask*) [11] e a IRM (*Ideal Reverberant Mask*) [12] utilizam o sinal limpo como referência. Já as máscaras cegas (*blind*), como a BRM (*Blind Reverberant Mask*) [13] e a BAM (*Blind Adaptive Mask*) [14], não requerem conhecimento prévio do sinal de voz.

Este estudo propõe a investigação do efeito das EN de sinais capturados por *drones*. As máscaras IBM, IRM e IRatioM são adotadas para aprimoramento da inteligibilidade dos sinais acústicos em diferentes cenários reverberantes e ruidosos. A avaliação dos resultados considera as medidas objetivas de inteligibilidade ESTOI (*Extended Short-Time Objective Intelligibility*) [15] e CSII (*Coherence and Speech Intelligibility Index*) [16]. Nos experimentos, sinais de voz da base TIMIT [17] são reverberados com a resposta ao impulso (RIR - Room Impulse Response) de duas salas reais: *LASP 2* da base LASP_RIR[1] e *Meeting* da base AIR [18]. Três EN - Decolagem, Motores e Voo - são adicionados aos sinais reverberados com valores de SNR de -10 dB, -5 dB, 0 dB e 5 dB. Adicionalmente, os ruídos ambientais Balbúrdia e Cafeteria também são incluídos nos experimentos. Os resultados demonstram que, após o processamento com as máscaras, os sinais corrompidos por EN obtiveram maior ganho de inteligibilidade quando comparados àqueles degradados por ruídos ambientais. Além disso, o método IRatioM obteve maior ganho de inteligibilidade para as condições analisadas.

V. Freitas é mestranda do Programa de Pós-Graduação em Engenharia em Engenharia Elétrica do Instituto Militar de Engenharia (IME). O trabalho de pesquisa é desenvolvido no Laboratório de Processamento de Sinais Acústicos (LASP/IME) e parcialmente financiado pelo CNPq (308155/2019-0) e pela FAPERJ (200518/2023). E-mails: viviane.freitas,coelho@ime.eb.br.

[1]Disponível em lasp.ime.eb.br.

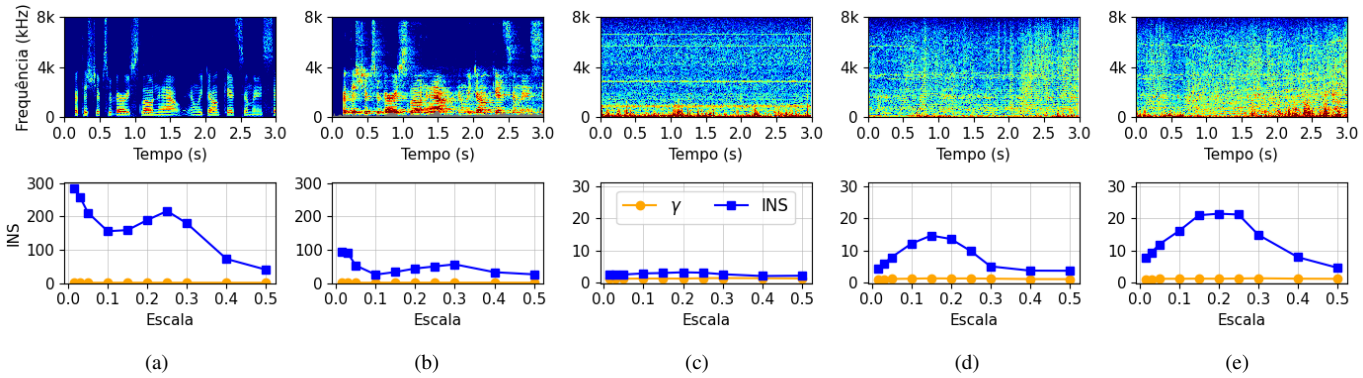


Fig. 2. Espectrogramas e respectivos valores de INS para: (a) Sinal de voz da base TIMIT [17]; (b) Sinal de voz reverberado; (c) EN 'Motores' (d) EN 'Decolagem' (e) EN 'Voo' da base DREGON [19].

O restante deste artigo está organizado da seguinte forma: A Seção II descreve os efeitos das ego-interferências de *drones* em sinais de voz. A Seção III apresenta os métodos de aplicação máscaras acústicas ideais estudados. Na Seção IV, avalia-se os resultados do processamento com diferentes máscaras para os cenários ruidosos. Por fim, a Seção V indica a conclusão.

## II. Ego-interferências e efeitos na inteligibilidade

As ego-interferências possuem características temporais e espectrais que variam de acordo com o movimento do *drone*. A Figura 2 mostra o espectrograma e o índice de não-estacionariedade (INS, *Index of nonstationarity*)[20] de um sinal de voz da base TIMIT, sua versão reverberada com RIR da sala *LASP 2*, e três EN [19]. O INS foi proposto como uma medida objetiva para avaliar a não-estacionariedade dos sinais. Para isso, o sinal analisado é comparado a referenciais estacionários, chamados *surrogates* [20]. Seja $T_h$ o tamanho da janela de análise espectral e $T$ a duração total do sinal, o valor de INS é computado para cada escala de observação $T_h/T$. Conforme proposto em [20], um limiar de estacionariedade é definido considerando precisão de 95% de forma que:

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

Note da Figura 2(a)-(b) que os sinais de voz são altamente não-estacionários, isto é, INS $\gg \gamma$. O valor máximo de INS (INS$_{max}$) é de 283,81 para o sinal de voz e 94,63 para sua versão reverberada. A Figura 2(d)-(e) mostra também que as EN Voo e Decolagem são não-estacionárias, visto que INS $> \gamma$ para todas as escalas de observação. Além disso, estas EN têm energia distribuída em toda a faixa de frequências $[0, 8]$ kHz. Já a EN Motores, Figura 2(c), é estacionária, com INS $\approx \gamma$ para a maioria das escalas.

Os efeitos das variações acústicas na inteligibilidade podem ser investigados a partir da análise de sinais ruidosos e reverberantes-ruidosos. A Figura 3 ilustra quatro cenários nos quais os sinais de voz da Figura 2(a)-(b) estão corrompidos pela EN Voo. A Figura 3(a)-(b) corresponde a SNR de 0 dB, enquanto a Figura 3(c)-(d) considera SNR de $-10$ dB. O sinal

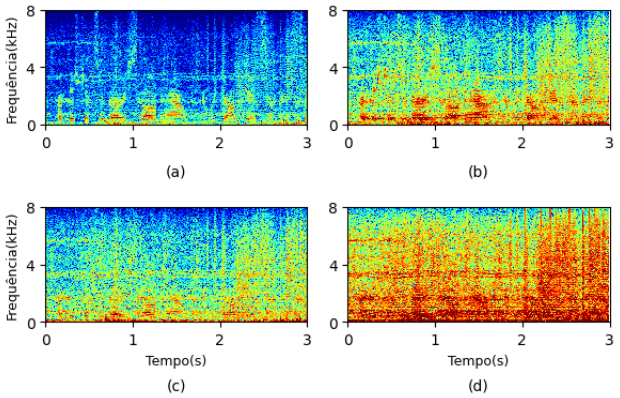


Fig. 3. (a) Sinal de voz corrompido com EN Voo com SNR=0 dB; (b) Sinal de voz corrompido com EN Voo com SNR=-10 dB; (c) Sinal de voz reverberado com RIR da sala *LASP 2* e corrompido com EN Voo com SNR=0 dB; (d) Sinal de voz reverberado com RIR da sala *LASP 2* e corrompido com EN Voo com SNR=-10 dB. [19].

corrompido pela EN com SNR de $-10$ dB corresponde a condições frequentemente enfrentadas durante o emprego de técnicas de *drone audition*. Note na Figura 3(d) que a presença simultânea da reverberação *LASP 2* e da EN Voo provoca interferências significativas nas faixas de frequências onde a voz está presente. A inteligibilidade diminui severamente nessa situação, à medida que aumentam os efeitos de ruído e reverberação.

## III. Máscaras ideais para aprimoramento da inteligibilidade acústica

Máscaras acústicas são métodos clássicos para aprimoramento da inteligibilidade de sinais de voz. Estes métodos são compostos pelas etapas de decomposição, cálculo de parâmetros, definição dos valores associados a cada componente, mascaramento e reconstrução do sinal. Métodos binários classificam cada região tempo-frequência (T-F) do sinal como composta principalmente pelo sinal (*speech-dominant*) ou pela interferência (*noise-dominant*). Os quadros classificados com predominância do sinal são preservados, enquanto os demais são descartados [14]. A seleção depende da comparação entre um ou mais parâmetros, calculados a partir de quadros T-F,

e limiares de decisão. Em máscaras não-binárias, ou suaves, como a *Ideal Ratio Mask* (IRatioM) [11], a etapa de classificação é substituída pelo cálculo de uma função peso que atenua os quadros com valores contínuos proporcionais a uma razão que caracteriza a presença das interferências.

Neste trabalho, a *Ideal Reverberant Mask* (IRM), *Ideal Binary Mask* (IBM) e a *Ideal Ratio Mask* (IRatioM) são empregadas para o processamento de sinais corrompidos por ruído e reverberação. O passo-a-passo para a definição de cada máscara é detalhado a seguir.

### A. Ideal Reverberant Mask (IRM)

A máscara IRM é determinada a partir da relação sinal-ruído reverberante (SRR) em cada unidade T-F. Após a decomposição, o SRR é obtido pela razão entre a energia do sinal limpo e a energia do sinal reverberante, como mostra a equação (2).

Para cada instante de tempo $t$ e banda de frequência $k$, a energia é calculada como a soma dos quadrados das amplitudes no domínio temporal dentro da janela correspondente àquela unidade e banda de frequência.

$$\text{SRR}(t,k) = 10\log_{10}\left(\frac{E_{\text{direto}}(t,k)}{E_{\text{reverberado}}(t,k)}\right). \tag{2}$$

O limiar $T_0$ utilizado para classificar cada unidade é fixado com base em testes experimentais em $-8$ dB [21]. A máscara é construída comparando a SRR de cada unidade com um limiar fixo $T_0$:

$$\text{IRM}(t,f) = \begin{cases} 1, & \text{se SRR}(t,k) > T_0, \\ 0, & \text{caso contrário.} \end{cases} \tag{3}$$

### B. Máscara IBM

A *Ideal Binary Mask* (IBM) é baseada em um processo de classificação binária. Calcula-se o valor de SNR considerando as componentes do sinal de voz e do ruído em cada unidade T-F. Quando a SNR em uma componente excede um limiar local (*Local Criterion*, LC), a unidade é considerada dominada pela voz e recebe valor 1; caso contrário, recebe valor 0. Desta forma,

$$\text{IBM}(t,k) = \begin{cases} 1, & \text{se SNR}(t,k) > \text{LC}, \\ 0, & \text{caso contrário.} \end{cases} \tag{4}$$

O limiar LC é comumente fixado em um valor inferior à SNR global do sinal. Neste trabalho, o LC utilizado foi de $-5$ dB [10].

### C. Máscara IRatioM

Apesar de produzir melhorias expressivas na inteligibilidade, a IBM impõe uma supressão abrupta das regiões mascaradas, o que pode introduzir artefatos perceptuais e degradar a qualidade sonora. Além disso, é importante observar que, devido ao limiar fixo, o aprimoramento da inteligibilidade pode ser comprometido quando a SNR for negativa.

A *Ideal Ratio Mask* (IRatioM) representa uma extensão suavizada da proposta de mascaramento no domínio tempo-frequência. Ao invés de valores binários, atribui-se valores contínuos entre 0 e 1 às unidades T-F da máscara. Isso proporciona uma atenuação progressiva das unidades com interferência, evitando a supressão abrupta das regiões mascaradas, o que pode introduzir artefatos perceptuais e degradar a qualidade sonora. Com isso, preserva-se as características do sinal de voz. A definição da máscara IRatioM é dada por:

$$\text{IRatioM}(t,k) = \sqrt{\frac{\text{SNR}(t,k)}{\text{SNR}(t,k)+1}}. \tag{5}$$

TABELA I

MÉDIA DA ESTOI DE SINAIS NÃO-PROCESSADOS (NP) EM DIFERENTES CONDIÇÕES DE EGO-INTERFERÊNCIA

| | Ruído | SNR | NR | *Meeting* | *LASP 2* |
|---|---|---|---|---|---|
| Ego-interferências | Motores $\text{INS}_{\text{max}} = 3,29$ | -10 dB | 0,28 | 0,23 | 0,21 |
| | | -5 dB | 0,38 | 0,32 | 0,28 |
| | | 0 dB | 0,49 | 0,41 | 0,34 |
| | | 5 dB | 0,61 | 0,50 | 0,38 |
| | Decolagem $\text{INS}_{\text{max}} = 14,65$ | -10 dB | 0,21 | 0,17 | 0,14 |
| | | -5 dB | 0,32 | 0,25 | 0,22 |
| | | 0 dB | 0,44 | 0,35 | 0,30 |
| | | 5 dB | 0,57 | 0,46 | 0,36 |
| | Voo $\text{INS}_{\text{max}} = 21,40$ | -10 dB | 0,16 | 0,14 | 0,12 |
| | | -5 dB | 0,27 | 0,22 | 0,19 |
| | | 0 dB | 0,40 | 0,33 | 0,26 |
| | | 5 dB | 0,53 | 0,43 | 0,33 |

TABELA II

MÉDIA DA ESTOI DE SINAIS RUIDOSOS DE CENÁRIOS AMBIENTAIS NÃO-PROCESSADOS (NP)

| | Ruído | SNR | NR | *Meeting* | *LASP 2* |
|---|---|---|---|---|---|
| Ruídos ambientais | Balbúrdia $\text{INS}_{\text{max}} = 40,21$ | -5 dB | 0,27 | 0,23 | 0,20 |
| | | 0 dB | 0,38 | 0,32 | 0,26 |
| | | 5 dB | 0,50 | 0,43 | 0,32 |
| | Cafeteria $\text{INS}_{\text{max}} = 25,75$ | -5 dB | 0,28 | 0,25 | 0,22 |
| | | 0 dB | 0,40 | 0,35 | 0,28 |
| | | 5 dB | 0,53 | 0,46 | 0,35 |

## IV. EXPERIMENTOS, RESULTADOS E DISCUSSÃO

Nesta Seção, as máscaras IBM, a IRM e IRatioM são comparadas em 15 conjuntos de interferências acústicas. Para isso, 20 sinais de voz provenientes da base TIMIT [17] são reverberados e corrompidos por EN e ruídos ambientais. Para reverberação, foram consideradas as salas *Meeting* [18] e *LASP 2* com valores de $T_{60}$ de 0,36 e 0,79 e DRR de 1,22 e -4,37, respectivamente. As EN Motores, Decolagem e Voo [19] e os ruídos ambientais Balbúrdia e Cafeteria [22] são adicionados aos dois grupos de sinais reverberados, além de um terceiro grupo de sinais não reverberados (NR). Todos os sinais ruidosos possuem valores de SNR de -5 dB, 0 dB e 5 dB. Para as EN, sinais com a SNR de -10 dB são também considerados nos experimentos, totalizando 1080 testes para cada máscara.

As Tabelas I e II apresentam os valores médios de ESTOI para os sinais não-processados corrompidos pelas EN e pelos

TABELA III
RESULTADOS DE ΔESTOI PARA SINAIS COM EGO-INTERFERÊNCIA PROCESSADOS COM MÁSCARAS ACÚSTICAS IDEAIS.

| | | **Ego-interferências** | | | | | | | | | | | | Ruídos ambientais | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Motores** | | | | **Decolagem** | | | | **Voo** | | | | Balbúrdia | | | Cafeteria | | |
| | SNR(dB) | -10 | -5 | 0 | 5 | -10 | -5 | 0 | 5 | -10 | -5 | 0 | 5 | -5 | 0 | 5 | -5 | 0 | 5 |
| NR | IRM | 0,28 | 0,28 | 0,27 | 0,23 | 0,30 | 0,31 | 0,28 | 0,23 | 0,28 | 0,33 | 0,31 | 0,26 | 0,34 | 0,34 | 0,30 | **0,35** | **0,35** | 0,28 |
| | IBM | 0,26 | 0,25 | 0,23 | 0,18 | 0,27 | 0,27 | 0,25 | 0,20 | 0,24 | 0,30 | 0,27 | 0,22 | **0,31** | 0,30 | 0,26 | **0,31** | **0,31** | 0,25 |
| | IRatioM | 0,42 | 0,35 | 0,28 | 0,21 | 0,49 | 0,41 | 0,33 | 0,25 | **0,54** | 0,46 | 0,38 | 0,28 | 0,45 | 0,37 | 0,29 | 0,44 | 0,36 | 0,27 |
| *Meeting* | IRM | 0,28 | 0,30 | 0,28 | 0,24 | 0,29 | 0,32 | 0,29 | 0,25 | 0,26 | 0,30 | 0,31 | 0,28 | **0,34** | 0,32 | 0,29 | **0,34** | **0,34** | 0,28 |
| | IBM | 0,22 | 0,21 | 0,18 | 0,13 | 0,23 | 0,22 | 0,19 | 0,14 | 0,23 | 0,24 | 0,22 | 0,17 | **0,27** | 0,23 | 0,18 | 0,26 | 0,23 | 0,16 |
| | IRatioM | 0,45 | 0,38 | 0,31 | 0,25 | 0,51 | 0,44 | 0,36 | 0,29 | **0,55** | 0,47 | 0,40 | 0,32 | 0,46 | 0,39 | 0,32 | 0,47 | 0,40 | 0,30 |
| *LASP 2* | IRM | 0,29 | 0,28 | 0,29 | 0,29 | 0,32 | 0,32 | 0,31 | 0,30 | 0,28 | **0,33** | **0,33** | 0,32 | 0,32 | 0,32 | 0,32 | 0,30 | 0,30 | 0,31 |
| | IBM | 0,14 | 0,10 | 0,08 | 0,06 | 0,16 | 0,14 | 0,09 | 0,06 | 0,16 | 0,14 | 0,11 | 0,08 | **0,17** | 0,14 | 0,10 | 0,15 | 0,11 | 0,09 |
| | IRatioM | 0,44 | 0,38 | 0,34 | 0,30 | 0,51 | 0,44 | 0,38 | 0,33 | **0,55** | 0,48 | 0,42 | 0,36 | 0,45 | 0,40 | 0,35 | 0,46 | 0,42 | 0,37 |

ruídos ambientais, respectivamente. Como esperado, os valores de inteligibilidade diminuem com os valores de SNR. No entanto, é possível notar que a variação de ESTOI para SNR distintas diminui na presença de reverberação. Por exemplo, o valor de ESTOI varia de $0,12$ para $0,33$ quando a SNR varia de $-10$ dB até $5$ dB para a EN Voo e a reverberação *LASP 2*. Já para sinais não reverberados, a variação de ESTOI é de $0,16$ para $0,53$ para esta mesma EN. Resultados similares podem ser verificados na Tabela II. Para o exemplo do ruído Cafeteria, a variação de ESTOI é de $0,25$ (de $0,28$ para $0,53$) para sinais não reverberados. No caso de sinais reverberados com a RIR da sala *LASP 2*, esta variação é de $0,13$ (de $0,22$ para $0,35$).

A Tabela III apresenta os resultados de ΔESTOI para as três máscaras nas três condições de reverberação. Os valores em negrito correspondem aos melhores resultados para cada uma das máscaras acústicas. Para a SNR de $-10$ dB, característica de sinais capturados por *drones*, a IRatioM alcançou ganhos médios de $0,51$ (Decolagem/*LASP 2*) e 0,55 (Voo/*LASP 2*). Estes valores são superiores àqueles obtidos para ruídos ambientais em condições menos severas (SNR de $-5$ dB): $0,45$ (Balbúrdia) e $0,46$ (Cafeteria). O ruído Motores, sendo estacionário ($\text{INS}_{max} = 3,29$), permite às máscaras binárias alcançarem desempenho mais próximo à IRatioM, enquanto os ruídos não-estacionários Decolagem ($\text{INS}_{max} = 14,65$) e Voo ($\text{INS}_{max} = 21,40$) beneficiam-se significativamente do mascaramento suave.

A Figura 4 ilustra as curvas do ganho médio de inteligibilidade (ΔESTOI) dos sinais processados pelas máscaras acústicas para cada SNR. O ganho de inteligibilidade é aqui calculado como a diferença entre os valores de ESTOI [15] do sinal processado pela máscara acústica e do sinal não processado. As curvas evidenciam como a máscara suave *IRatioM* tem resultados de inteligibilidade objetiva significativamente melhores que as máscaras binárias, principalmente para SNR menor que 0 dB.

Note da Tabela III e da Figura 4 que o aprimoramento da inteligibilidade pela máscara IBM diminui à medida que se intensificam os efeitos da reverberação. Na sala *LASP 2* ($T_{60} = 0,79$ s), os ganhos da IBM para EN em SNR $-10$ dB variam entre $0,14$-$0,16$, contrastando com $0,22$-$0,27$ na sala *Meeting* ($T_{60} = 0,36$ s). Esta degradação é menos pronunciada para a IRM, cujos resultados sugerem que o aumento da $T_{60}$ não prejudica significativamente a ESTOI dos sinais processados, mantendo ganhos consistentes entre 0,26-0,32 independentemente da reverberação. O ganho médio de inteligibilidade observado na Figura 4 é diferente cada EN. Para a máscara IRatioM o valor de ΔESTOI é maior para as EN com maior valor de $\text{INS}_{max}$.

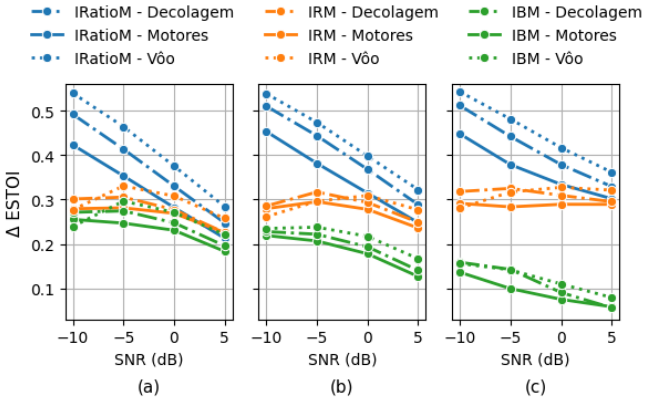


Fig. 4. Ganho médio de inteligibilidade (ΔESTOI) para diferentes máscaras acústicas e EN, considerando (a) sinais de voz não reverberados; (b) sinais reverberados com a RIR da sala *Meeting*; e (c) sinais reverberados com a RIR da sala *LASP 2*.

A Figura 5 mostra os resultados de ΔESTOI para sinais corrompidos por ruídos ambientais. Novamente, os melhores resultados são obtidos pela máscara IRatioM. Além disso, os resultados da IBM, assim como na Figura 4, são degradados para os sinais reverberantes. No entanto, é possível observar que as máscaras binárias IBM e IRM resultam em ganhos no valor de ESTOI superiores quando comparados ao aprimoramento obtido para as EN. Em contrapartida, a IRatioM é mais eficaz ao tratar as EN em todos os cenários analisados.

A Figura 6 exibe o *boxplot* dos valores de CSII considerando sinais ruidosos nos cenários com reverberação e sem

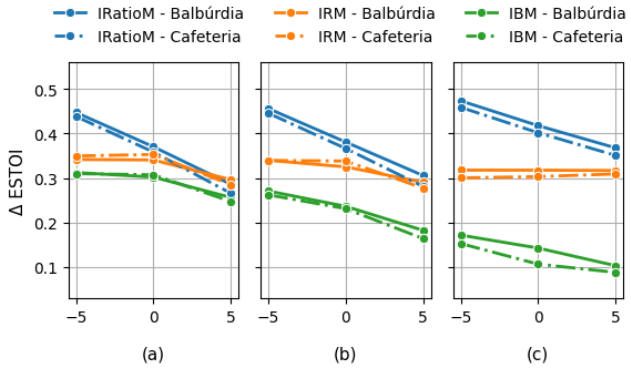


Fig. 5. Ganho médio de inteligibilidade (ΔESTOI) para diferentes máscaras acústicas e ruídos ambientais, considerando (a) sinais de voz não reverberados; (b) sinais reverberados com a RIR da sala *Meeting*; e (c) sinais reverberados com a RIR da sala *LASP 2*.

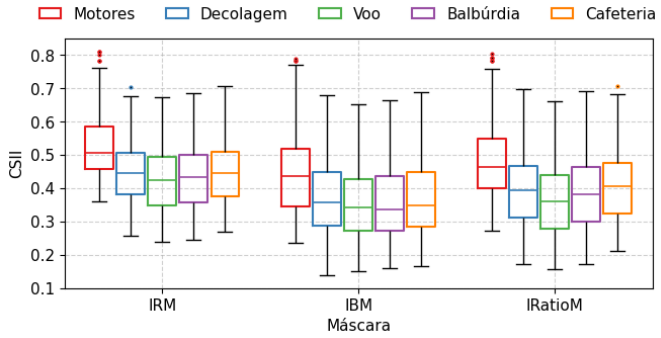


Fig. 6. *Boxplot* dos valores da medida CSII para os sinais ruidosos processados com diferentes máscaras acústicas.

reverberação. Neste caso, foram considerados sinais com SNR de $-5$ dB, 0 dB e 5 dB. Observe que a maior inteligibilidade é alcançada para a EN estacionária Motores. Por outro lado, os menores resultados de CSII são obtidos para a EN não-estacionária Voo. Finalmente, é importante ressaltar que a máscara binária IRM alcançou os maiores resultados absolutos de CSII.

## V. Conclusão

Este estudo investigou os efeitos de ego-interferências de *drones* na inteligibilidade de sinais de voz ruidosos e processados com três máscaras acústicas ideais. Os cenários consideraram reverberação com RIR de duas salas reais, além de sinais não reverberados. Três EN com distintas características de não-estacionariedade e dois ruídos ambientais foram utilizados para corromper os sinais de voz.

Resultados da medida ESTOI mostraram que a máscara suave IRatioM obteve os maiores ganhos médios de inteligibilidade para todas as condições reverberantes-ruidosas. Em particular, os maiores valores de ΔESTOI foram alcançados para SNR de $-10$ dB, que mais se aproxima da condição real em *drone audition*. Além disso, o aprimoramento de ESTOI com a máscara IBM diminuiu conforme aumentaram os efeitos de reverberação. Em termos de CSII, os melhores resultados para as três máscaras foram obtidos para a EN estacionária Motores. Em geral, a máscara IRM atingiu os maiores índices de inteligibilidade para todas as EN e ruídos ambientais. A mediana de CSII para a IRM variou de 0,423 para a EN Voo até 0,507 para a EN Motores.

## Referências


[1] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors Journal*, vol. 17, no. 8, pp. 2447–2455, 2017.

[2] J. Martinez-Carranza and C. Rascon, "A review on auditory perception for unmanned aerial vehicles," *Sensors*, vol. 20, no. 24, 2020.

[3] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 152–158, 2016.

[4] L. Wang and A. Cavallaro, "Deep-learning-assisted sound source localization from a flying drone," *IEEE Sensors Journal*, vol. 22, no. 21, pp. 20828–20838, 2022.

[5] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2523–2537, 2020.

[6] G. Zucatelli and R. Coelho, "Adaptive reverberation absorption using non-stationary masking components detection for intelligibility improvement," *IEEE Signal Processing Letters*, vol. 27, pp. 1–5, 2020.

[7] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, pp. 181–197, 2005.

[8] G. Kim and P. C. Loizou, "Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints," *IEEE Signal Processing Letters*, vol. 17, no. 12, pp. 1010–1013, 2010.

[9] L. Zão, D. Cavalcante, and R. Coelho, "Time-frequency feature and ams-gmm mask for acoustic emotion classification," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 620–624, 2014.

[10] D. Wang, U. Kjems, M. Pedersen, J. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *The Journal of the Acoustical Society of America*, vol. 125, pp. 2336–47, 05 2009.

[11] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, pp. 1486–1501, 11 2006.

[12] N. Li and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1673–82, 04 2008.

[13] O. Hazrati, J. Lee, and P. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 133, pp. 1607–14, 03 2013.

[14] F. Farias and R. Coelho, "Blind adaptive mask to improve intelligibility of non-stationary noisy speech," *IEEE Signal Processing Letters*, vol. 28, pp. 1170–1174, 2021.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[16] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, pp. 2224–37, 05 2005.

[17] J. Garofolo, L. Lamel, W. Fischer, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," in *Linguist. Data Consortium*, Philadelphia, PA, USA, 1993.

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

[19] M. Strauß, P. Mordel, V. Miguet, and A. Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," pp. 1–8, 10 2018.

[20] 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, no. 7, pp. 3459–3470, 2010.

[21] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *Journal of the Acoustical Society of America*, vol. 129, pp. 3221–3232, May 2011.

[22] H. J. Steeneken and F. W. Geurtsen, "Description of the rsg-10 noise-database," *report IZF*, vol. 3, 1988.