Avaliação da Predição Objetiva da Inteligibilidade de Sinais Reverberantes e Ruidosos com Uso de Máscaras Acústicas

R. Martiny, R. Alcântara e R. Coelho

***Resumo*— Este artigo apresenta um estudo do impacto de ruído e reverberação na inteligibilidade e qualidade de sinais de voz. Esta avaliação inclui a utilização das máscaras acústicas ideais IBM, IRM e IQM, além das não-ideais BRM e BAM. As medidas objetivas de inteligibilidade ESII e STOI e de qualidade PEAQ e PESQ foram consideradas nos experimentos. Também foi analisado o efeito do ruído e da reverberação em relação à estacionariedade do sinal de voz. Os resultado mostraram que as máscaras ideais melhoraram a inteligibilidade e qualidade nos cenários avaliados. Além disso, as máscaras cegas foram capazes de melhorar a inteligibilidade e qualidade em diversos cenários.**

***Palavras-Chave*— Inteligibilidade, Máscara acústica, Não-estacionariedade.**

***Abstract*— This article presents a study of the impact of noise and reverberation on speech intelligibility and quality. This evaluation includes the use of ideal masks IBM, IRM and IQM, and the blind masks BRM and BAM. The objective intelligibility measures ESII and STOI and quality measures PEAQ and PESQ are considered in the experiments. The impact of noise and reverberation on the speeech signal stationarity is also analyzed. The results show that the ideal masks are capable of improving the speech intelligibility and quality in the scenarios. In addition, blind masks were able to improve intelligibility and quality in several scenarios.**

***Keywords*— Intelligibility, Binary Mask, Non-Stationarity.**

I. Introdução

O sinal de voz é geralmente capturado em ambientes naturais com a presença de múltiplas variações ou interferências acústicas, tais como reverberação e ruídos. Estes efeitos afetam a inteligibilidade e podem fazer com que o ouvinte não capte toda a informação da fonte sonora de referência [1]. Além disso, aplicações de reconhecimento de palavras e de locutor têm a sua acurácia prejudicada nestas situações. Em particular, ruídos ambientais e reverberação impactam negativamente a audição de usuários de próteses auditivas e implantes cocleares [2].

No cenário do "*cocktail party*" [3], um ouvinte é capaz de selecionar e compreender uma única fonte sonora em meio a diversas interferências. Baseadas nesta característica do sistema auditivo, as máscaras acústicas [4] foram propostas inicialmente para aprimorar a inteligibilidade do sinal de voz em ambientes ruidosos. A técnica realiza uma decomposição em quadros tempo-frequência (TF) e atua em cada quadro de acordo com a sua dominância pela interferência. Máscaras acústicas podem ser classificadas entre ideais, que utilizam informações do sinal limpo, e não-ideais (cegas), que não utilizam conhecimento prévio do sinal, sendo mais adaptadas ao uso prático. Máscaras ideais apresentam melhor desempenho e possuem aplicação no treinamento de redes neurais [5]. Em [6], foi proposta a máscara acústica ideal IBM (*Ideal Binary Mask*). Nela, os valores de razão sinal-ruído (SNR - *signal-to-noise ratio*) são utilizados como critério para decidir se um quadro tempo-frequência será excluído do sinal com interferência. A máscara IRM (*Ideal Ratio Mask*) [7] define valores fracionários de ganho em cada quadro TF de acordo com a razão entre a energia do sinal limpo e a do sinal corrompido. Baseada na IRM, a máscara ideal IQM (*Ideal Quantized Mask*) utiliza valores discretos de níveis de atenuação [8]. Entre as máscaras cegas, a BRM (*Binary Reverberant Mask*) apresentou melhora de inteligibilidade para ambientes com reverberação em testes subjetivos [9]. Em cenários com ruídos ambientais, a máscara cega BAM (*Blind Adaptive Mask*) [10] obteve ganho de inteligibilidade e qualidade.

Este trabalho apresenta um estudo sobre o impacto da inteligibilidade e qualidade do sinal de voz causado por ambientes reverberantes e ruidosos. Em seguida, o desempenho de máscaras acústicas ideais e não-ideais é analisado nestas situações. Além disso, o efeito do ruído e da reverberação em relação à não-estacionariedade do sinal de voz é avaliado de acordo com o índice de não-estacionariedade (INS - *Index of Non-Stationarity*) [11]. Para a avaliação da inteligibilidade são utilizadas as medidas objetivas ESII (*Extended Speech Intelligibility Index*) [12] e STOI (*Short-Time Objective Intelligibility*) [13]. As medidas PEAQ (*Perceptual Evaluation of Audio Quality*) [14] e PESQ (*Perceptual Evaluation of Speech Quality*) [15] são usadas nos testes de qualidade do sinal de voz.

O restante deste artigo está organizado da seguinte maneira: Na Seção II são descritas as máscaras acústicas. A Seção III aborda as medidas de inteligibilidade e qualidade utilizadas para uma avaliação objetiva dos métodos. A descrição dos cenários experimentais e as discussões dos resultados são apresentadas na Seção IV. Por fim, na Seção V são apresentadas as principais conclusões deste trabalho.

II. Máscaras Acústicas Ideais e Cegas

Nesta Seção é apresentada uma breve descrição das máscaras acústicas utilizadas neste artigo. O objetivo do emprego das

R. Martiny é mestrando do Programa de Pós-Graduação da Engenharia Elétrica do Instituto Militar de Engenharia (IME); R. Alcântara é doutorando do Programa de Pós-Graduação da Engenharia de Defesa do IME. O trabalho dos autores R. Martiny, R. Alcântara e R. Coelho é desenvolvido no Laboratório de Processamento de Sinais Acústicos (LASP/IME), Rio de Janeiro, Brasil. E-mails: {rafaelmarinati,raoni,coelho}@ime.eb.br.

máscaras acústicas é a redução dos efeitos das interferências do ruído e da reverberação no sinal de voz, aprimorando a sua inteligibilidade e qualidade.

A. Decomposição tempo-frequência e reconstrução do sinal

As máscaras acústicas ideais e não-ideais geralmente utilizam decomposições em tempo-frequência. Sejam $x(t)$ e $y(t)$ o sinal limpo e corrompido por ruído e reverberação, respectivamente. Um banco de filtros *gammatone* [16] [17] é utilizado para dividir os sinais em sub-bandas. Em seguida, cada uma destas bandas é janelada, obtendo-se as componentes $Y(k,q)$, que representam a k-ésima sub-banda e o q-ésimo quadro do sinal corrompido. No caso das máscaras ideais, o mesmo procedimento é realizado com o sinal limpo para a obtenção de $X(k,q)$. Após a aplicação das máscaras, são obtidos quadros tempo-frequência $\hat{X}(k,q)$. A síntese do sinal mascarado é realizada através da concatenação destes quadros em cada sub-banda de acordo com a sua sobreposição original. Por fim, o sinal processado é obtido após a soma das sub-bandas reconstruídas.

B. Máscaras Acústicas Ideais

As máscaras acústicas ideais são definidas em cada componente TF a partir de $X(k,q)$ e $Y(k,q)$. As máscaras IBM, IRM e IQM estão detalhadas a seguir:

1) IBM: Na máscara IBM [6], o critério de seleção empregado é a SNR. Desta forma, o valor da máscara IBM(k,q) em cada componente TF é definido como:

$$\text{IBM}(k,q) = \begin{cases} 1, & \text{se SNR}(k,q) \geq \theta, \\ 0, & \text{caso contrário.} \end{cases} \tag{1}$$

sendo, a SNR(k,q) o valor da SNR na k-ésima sub-banda e q-ésimo quadro. As componentes do sinal mascarado são calculadas a partir de $\hat{X}(k,q) = \text{IBM}(k,q)Y(k,q)$.

2) IRM: Em [7], o sinal mascarado é resultado de um ganho que, para cada quadro tempo-frequência é dado por:

$$\text{IRM}(k,q) = \sqrt{\frac{S(k,q)}{S(k,q)+I(k,q)}}, \tag{2}$$

onde $S(k,q)$ e $I(k,q)$ são, respectivamente, a energia do sinal limpo, e da interferência do ruído e reverberação na k-ésima sub-banda e q-ésimo quadro. Por fim, os quadros tempo-frequência do sinal mascarado são obtidos de acordo com $\hat{X}(k,q) = \text{IRM}(k,q)Y(k,q)$.

3) IQM: A máscara IQM [8] utiliza IRM(k,q) como base para a obtenção de N possíveis valores de IQM(k,q). Isto é feito a partir de uma seleção de pontos da máscara IRM de acordo com:

$$p = -\log_2 \sqrt{\frac{10^{(\lambda/10)}}{10^{(\lambda/10)}+1}}, \tag{3}$$

$$h_n = \left(\frac{n-1}{N}\right)^p, \tag{4}$$

sendo, p definido a partir de um limiar λ, N o número total de níveis definidos na IQM e $n = 1, ..., N$. Os valores de $\text{IQM}_N(k,q)$ são calculados por:

$$\text{IQM}_N(k,q) = \begin{cases} h_1, \text{se } 0 \leq \text{IRM}(k,q) \leq h_2 \\ h_2, \text{se } h_2 < \text{IRM}(k,q) \leq h_3 \\ \vdots \\ h_N, \text{se } h_N < \text{IRM}(k,q) \leq 1. \end{cases} \tag{5}$$

Por fim, os quadros tempo-frequência do sinal mascarado são obtidos a partir de $\hat{X}(k,q) = \text{IQM}_N(k,q)Y(k,q)$.

C. Máscaras Acústicas Cegas

As máscaras cegas (não-ideais) são mais adaptadas ao uso prático que as ideais por não terem a limitação de necessitar informações do sinal de voz limpo. Neste estudo, foram utilizadas as máscaras não-ideais BRM e BAM, descritas abaixo:

1) BRM: A máscara BRM [9] foi proposta para supressão do efeito da reverberação. Para cada unidade tempo-frequência é calculado um coeficiente dado por:

$$f_M(k,q) = 10\log_{10}\left(\frac{\sigma^2_{r'}(k,q)}{\sigma^2_{|r|}(k,q)}\right) \tag{6}$$

onde $r'(k,q) = |r(k,q)|^\alpha$, sendo $|r(k,q)|$ o valor absoluto das amostras no quadro *q* e sub-banda *k*. Os valores de f_M são suavizados através de um filtro mediana de ordem 3.

O critério de seleção binário é baseado no histograma $f_{\text{hist}}(k,q)$, calculado a partir dos valores de f_M dos Q_a quadros anteriores a *t* e Q_p quadros posteriores. Cada histograma possui L classes com pesos p_i $(i = 1, 2, ..., L)$. Com estes valores, são calculadas a média global m_G, a média cumulativa $m(l)$ e a soma cumulativa $P_s(l)$, definidas por:

$$m_G = \sum_{i=1}^{L} i.p_i \qquad m(l) = \sum_{i=1}^{l} i.p_i \qquad P_s(l) = \sum_{i=1}^{l} p_i. \tag{7}$$

O limiar ótimo l^* é obtido pelo índice l que maximiza a variância entre classes $\sigma^2_B(L)$, dada por:

$$\sigma^2_B(l) = \frac{(m_G P_s(l) - m(l))^2}{P_s(l)(1-P_s(l))}. \tag{8}$$

Em seguida, a máscara BRM é dada por:

$$\text{BRM}(k,q) = \begin{cases} 1, & \text{se } f_M(k,q) \geq \max(l^*(k,q), l_0), \\ 0, & \text{caso contrário,} \end{cases} \tag{9}$$

onde l_0 é um limiar definido para o silêncio. Por fim, as componentes do sinal mascarado são calculadas através de $\hat{X}(k,q) = \text{BRM}(k,q)Y(k,q)$.

2) BAM: A máscara BAM [10] não utiliza decomposições em sub-bandas. O sinal é dividido em quadros no domínio do tempo. Depois, são estimadas as componentes de ruído e a proporção do sinal de voz em cada um desses quadros através estimador DATE (*d-Dimensional Trimmed Estimator*) [18].

O parâmetro d_q é definido para identificar quadros onde o sinal alvo é predominante, dado por:

$$d_q = \frac{|\sigma_{q_{ny}} - \hat{\sigma}_q|}{|\sigma_{q_{ny}} + \hat{\sigma_q}|}, \tag{10}$$

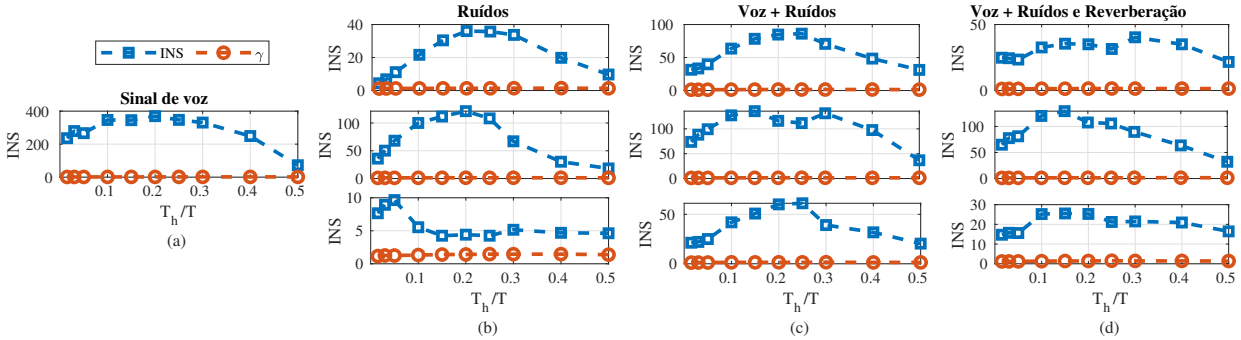

Fig. 1. INS para (a) Sinal limpo, (b) Ruídos, de cima para baixo, Balbúrdia, Motosserra e Fábrica, (c) Voz adicionadas dos ruídos Balbúrdia, Motosserra e Fábrica com SNR = 0 dB e (d) Voz com reverberação Escadaria adicionada dos ruídos Balbúrdia, Motosserra e Fábrica com SNR = 0 dB.

TABELA I

REVERBERAÇÕES SELECIONADAS DAS BASES DE DADOS LASP E AIR.

Reverberação	RT_{60} (s)	d_{fm} (m)	DRR (dB)
LASP	0,79	1,00	-4,37
Escadaria	1,10	1,00	-5,27

sendo, $\hat{\sigma_q}$ o desvio padrão do quadro q do ruído estimado pelo DATE e $\sigma_{q_{ny}}$ o desvio padrão do quadro q do sinal ruidoso.

A partir desses valores, a máscara define quais quadros permanecerão inalterados. Os restantes terão a sua amplitude reduzida através de um fator de subtração proporcional ao desvio padrão das suas amostras.

III. MEDIDAS OBJETIVAS DE INTELIGIBILIDADE E QUALIDADE

Nesta Seção estão descritas as medidas objetivas de inteligibilidade (ESII e STOI) e de qualidade (PEAQ e PESQ) utilizadas neste estudo.

A. ESII

A medida ESII [19] foi desenvolvida a partir da SII [20] para ser capaz de lidar com o comportamento não-estacionário das distorções. Para isso, a medida utiliza divisão do sinal de voz e do ruído em quadros. Em seguida, aplica-se a medida SII para cada um desses quadros e é calculada a média destes valores. Além disso, a ESII atribui relevâncias distintas para cada banda crítica do sinal de voz, adotando pesos atribuídos aos padrões da audição humana [21].

B. STOI

A medida de inteligibilidade STOI [13] é adequada para diversos tipos de degradações e possui alta correlação (ρ = 0,95) com testes subjetivos de inteligibilidade. Nela, são realizadas segmentações e decomposições em TF de ambos os sinais, limpo e processado. Em seguida, calcula-se o índice STOI intermediário, definido como uma estimativa do coeficiente de correlação linear entre as unidades TF desses sinais. Por fim, o índice STOI é dado pela média de todos os valores de STOI intermediário entre quadros e sub-bandas.

C. PEAQ

Essa medida de qualidade é baseada em atributos perceptuais denominados MOV (*Model Output Variables*). Estes atributos são comumente computados estimando limiares de mascaramento de um sinal de erro ou comparando as representações auditivas entre sinais de referência e processados. Por fim, os valores MOV são combinados de modo que resulte em uma única saída ODG (*Objective Difference Grade*) [14].

TABELA II

INS MÁXIMO DOS SINAIS DE VOZ REVERBERANTES E RUIDOSOS.

		LASP			Escadaria		
SNR (dB)		-5	0	5	-5	0	5
Ruídos	Balbúrdia	43,56	70,02	113,38	33,37	40,33	65,20
	Motosserra	190,14	217,63	255,48	115,16	129,44	158,30
	Fábrica	12,71	26,69	49,13	15,34	25,54	49,13

D. PESQ

A medida de qualidade PESQ [15] foi inicialmente proposta para examinar a qualidade dos sinais de voz em banda estreita ou telefonia celular. Primeiramente é utilizada uma FFT (*Fast Fourier Transform*) no sinal limpo e no distorcido para modelar um padrão telefônico. Em seguida, os sinais são alinhados no tempo e processados por uma transformação auditiva com frequência perceptiva na escala de Bark. Por fim, a medida PESQ é convertida para a escala MOS (*mean opinion score*), cujos valores variam de -0,5 (ruim) a 4,5 (sem distorção).

IV. EXPERIMENTOS: RESULTADOS E DISCUSSÃO

Experimentos foram realizados para a avaliação da inteligibilidade e qualidade objetiva dos sinais de voz reverberantes e ruidosos antes e após a utilização das máscaras acústicas IBM, IRM, IQM, BRM e BAM. Um subconjunto de 24 locutores (16 masculinos e 8 femininos) selecionados aleatoriamente da base de voz TIMIT [22] foi utilizado nos experimentos. No total, são utilizados 240 sinais de voz amostrados a 16 kHz e com duração média de 4 segundos. As reverberações foram extraídas da sala LASP[1] e Escadaria, obtida da base de dados AIR [23]. As duas reverberações foram captadas com distância fonte-microfone (d_{fm}) de 1 m e possuem diferentes valores de RT_{60}[2] e DRR (*Direct-to-Reverberant Ratio*), conforme apresenta a Tabela I. Após a convolução dos sinais de voz com as respostas ao impulso das salas, foram adicionados os ruídos acústicos Balbúrdia e Fábrica selecionados da base de dados RSG-10 [24] e o ruído Motosserra, de Freesound.org[3]. Os valores de SNR entre os sinais reverberados e os ruídos foram de -5 dB, 0 dB e 5 dB. A decomposição em sub-bandas nas máscaras acústicas foi realizada com um banco de 64 filtros. O janelamento foi realizado com quadros de 20 ms e 50% de sobreposição.

[1]Disponível em: http:lasp.ime.eb.br

[2]*Reverberation time*: Tempo necessário para que a energia da RIR decaia em 60 dB.

[3]Disponível em: www.freesound.org.

TABELA III

RESULTADOS DE ESII PARA SINAIS NÃO PROCESSADOS COM REVERBERAÇÃO E RUÍDO.

		LASP			Escadaria		
SNR (dB)		-5	0	5	-5	0	5
Ruídos	Balbúrdia	0,19	0,26	0,32	0,16	0,23	0,30
	Motosserra	0,15	0,21	0,28	0,15	0,21	0,28
	Fábrica	0,13	0,20	0,26	*0,12*	0,18	0,25

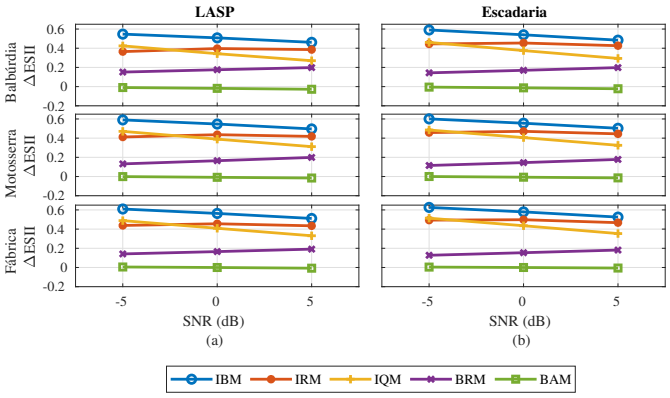

Fig. 2. Resultados de ΔESII para as salas (a) LASP e (b) Escadaria.

TABELA IV

RESULTADOS DE STOI PARA DIFERENTES RUÍDOS E SALAS.

		LASP				Escadaria			
SNR (dB)		-5	0	5	Média	-5	0	5	Média
Balbúrdia	NP	0,28	0,39	0,48	0,39	0,24	0,37	0,49	0,37
	IBM	0,74	0,77	0,80	0,77	0,77	0,81	0,84	0,81
	IRM	**0,83**	**0,84**	**0,85**	**0,84**	**0,84**	**0,85**	**0,86**	**0,85**
	IQM	0,80	0,83	0,84	0,82	0,83	**0,85**	**0,86**	**0,85**
	BRM	**0,25**	**0,34**	**0,44**	**0,34**	**0,28**	**0,39**	**0,50**	**0,39**
	BAM	0,16	0,29	0,42	0,29	0,23	0,37	0,50	0,36
Motosserra	NP	0,29	0,40	0,50	0,39	0,25	0,37	0,49	0,37
	IBM	0,72	0,76	0,79	0,76	0,73	0,79	0,83	0,79
	IRM	**0,84**	**0,84**	**0,85**	**0,84**	**0,84**	**0,86**	**0,87**	**0,86**
	IQM	0,80	0,83	**0,85**	0,82	0,82	0,85	**0,87**	0,85
	BRM	**0,25**	**0,34**	**0,45**	**0,35**	**0,29**	**0,39**	**0,51**	**0,40**
	BAM	0,18	0,29	0,42	0,29	0,23	0,36	0,50	0,36
Fábrica	NP	0,26	0,38	0,48	0,37	0,24	0,36	0,48	0,36
	IBM	0,69	0,75	0,79	0,74	0,71	0,78	0,83	0,78
	IRM	**0,83**	**0,84**	**0,85**	**0,84**	**0,84**	**0,85**	**0,86**	**0,85**
	IQM	0,80	0,82	0,84	0,82	0,83	**0,85**	**0,86**	**0,85**
	BRM	**0,20**	**0,30**	**0,42**	**0,31**	**0,27**	**0,37**	**0,49**	**0,38**
	BAM	0,15	0,28	0,41	0,28	0,23	0,36	**0,49**	0,36

O INS (*Index of Non-Stationarity*) [11] foi adotado como medida para examinar a não-estacionariedade dos sinais em diferentes condições. Essa medida compara o sinal de voz com referências estacionárias, chamadas de *surrogates*. Além disso, utiliza-se uma escala T_h/T, que é a razão entre o tamanho da janela (T_h) e o tamanho total do sinal (T). Um limiar γ de estacionariedade é definido com 95% de precisão. Os ruídos são não-estacionários quando o INS é maior do que esse limiar. A Figura 1 mostra as curvas de INS para um sinal de voz, para os ruídos selecionados nos experimentos, para os sinal de voz corrompido pelos ruídos e para a voz com ruído e reverberação. A curva de INS do sinal de voz está acima do limiar, sendo considerada não-estacionária e com INS máximo de 369,71. As curvas em (b) indicaram que os três ruídos são não-estacionários, com Balbúrdia apresentando INS máximo de 36,02 e Fábrica de 9,65. O ruído com maior INS foi o Motosserra, com INS de 121,39. Em (c), as curvas mostraram que a adição dos ruídos diminuiu a não-estacionariedade do sinal de voz, reduzindo o índice. Por fim, a presença de reverberação, nas curvas em (d) causaram maior redução do INS. Na Tabela II os valores de INS são complementados para as SNR de -5 e 5 dB, e para a reverberação LASP. Nela, percebe-se a influência da sala no INS, mostrando que a sala Escadaria, de maior RT_{60}, tornou o sinal reverberado mais estacionário.

A. Resultados de inteligibilidade

1) ESII: A Tabela III apresenta os resultados de ESII para os sinais de voz reverberantes e ruidosos. Os resultados mostraram que a medida ESII apontou maior impacto na inteligibilidade para o ruído Fábrica. Além disso, mostrou que os sinais com reverberação Escadaria, com maior RT_{60}, tiveram menor inteligibilidade. Na Figura 2 está ilustrada a curva com os valores de ganho do índice ESII (ΔESII) dos sinais processados pelas máscaras acústicas para cada SNR. Nela, nota-se que a máscara ideal IBM obteve maiores ganhos em inteligibilidade. Além disso, alcançou uma média geral de 0,76, enquanto a IRM e IQM obtiveram 0,66 e 0,61, respectivamente. Com relação às máscaras não-ideais, os resultados mostraram que a BRM obteve melhor desempenho. Perceba que alcançou uma média de 0,48 para o Balbúrdia na sala LASP, enquanto a BAM obteve 0,24.

2) STOI: A Tabela IV apresenta os resultados de STOI para os sinais reverberantes e ruidosos. Perceba que os experimentos não processados com o ambiente LASP apresentam melhores resultados do que aqueles com a reverberação Escadaria. Isso significa que, para os experimentos, existe uma relação inversa de inteligibilidade entre o INS do sinal e o RT_{60} da sala. Além disso, é possível notar que a IRM apresentou maior índice STOI, com valor de aumento médio do índice de 0,49 para Escadaria e 0,46 para a LASP. A máscara não-ideal BRM melhorou a inteligibilidade em Escadaria, com aumento médio do índice STOI em 0,02, tendo melhor desempenho em -5 dB.

B. Resultados de qualidade

1) PEAQ: Os resultados de qualidade obtidos com PEAQ estão apresentados em forma de *Boxplot* na Figura 3 agrupados de acordo com os valores de SNR -5 dB, 0 dB e 5 dB. Os gráficos mostram que, entre as máscaras ideais, a IRM alcançou melhores ganhos de qualidade em todos os cenários, seguida da IQM e IBM. Entre as máscaras não-ideais, a BRM obteve maiores valores de PEAQ.

2) PESQ: A Tabela V ressalta os valores de qualidade obtidos pela medida PESQ. Nota-se a máscara IRM apresentou os melhores resultados para todas os ruídos e reverberações, com aumento em média de até 2,10 para o ruído Fábrica e reverberação Escadaria. A máscara BAM apresentou melhores resultados entre as não-ideais, com aumento de até 0,09 com o ruído Motosserra e reverberação Escadaria.

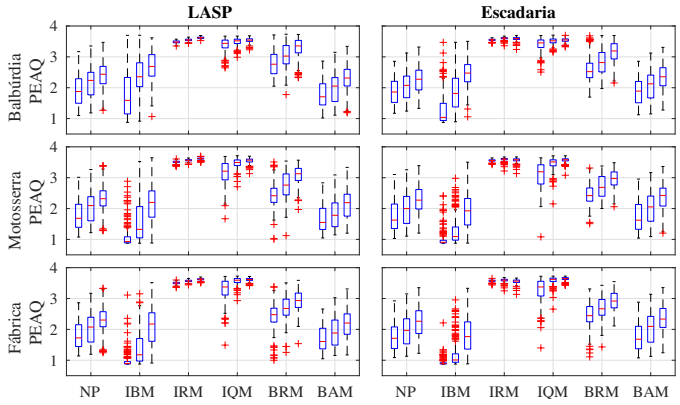

Fig. 3. Resultados de PEAQ para os experimentos com ruído e reverberação. Os resultados estão agrupados, da esquerda para a direita, de acordo com os valores de SNR -5 dB, 0 dB e 5 dB.

TABELA V

RESULTADOS DE PESQ PARA DIFERENTES RUÍDOS E SALAS.

		LASP				Escadaria			
	SNR (dB)	-5	0	5	Média	-5	0	5	Média
Balbúrdia	NP	0,39	0,65	1,01	0,68	0,54	0,74	1,12	0,80
	IBM	0,84	1,19	1,50	1,18	1,26	1,59	1,86	1,57
	IRM	**2,53**	**2,67**	**2,79**	**2,66**	**2,69**	**2,83**	**2,95**	**2,82**
	IQM	1,73	1,99	2,23	1,98	2,01	2,27	2,49	2,26
	BRM	0,37	0,24	0,42	0,34	0,37	0,47	0,71	0,52
	BAM	**0,64**	**0,55**	**0,78**	**0,65**	**0,63**	**0,82**	**1,20**	**0,88**
Motosserra	NP	0,61	0,80	1,13	0,85	0,69	0,89	1,18	0,92
	IBM	0,76	1,13	1,46	1,11	1,21	1,52	1,80	1,51
	IRM	**2,44**	**2,60**	**2,75**	**2,60**	**2,63**	**2,80**	**2,93**	**2,79**
	IQM	1,67	1,95	2,20	1,94	1,96	2,22	2,46	2,22
	BRM	0,36	0,37	0,61	0,45	0,49	0,64	0,94	0,69
	BAM	**0,64**	**0,72**	**0,91**	**0,76**	**0,81**	**0,93**	**1,28**	**1,01**
Fábrica	NP	0,31	0,61	0,95	0,62	0,41	0,75	1,14	0,77
	IBM	0,70	1,09	1,44	1,07	1,09	1,44	1,74	1,43
	IRM	**2,59**	**2,72**	**2,84**	**2,71**	**2,74**	**2,88**	**3,00**	**2,87**
	IQM	1,61	1,92	2,20	1,91	1,89	2,18	2,45	2,17
	BRM	0,12	0,21	0,41	0,25	0,25	0,38	0,50	0,38
	BAM	**0,19**	**0,41**	**0,74**	**0,45**	**0,46**	**0,81**	**1,24**	**0,84**

V. CONCLUSÃO

Este artigo apresentou um estudo da inteligibilidade e qualidade para sinais reverberantes e ruidosos e o desempenho de máscaras acústicas ideais e não-ideais em recuperar estas características. Os resultados mostraram que as máscaras acústicas ideais foram capazes de melhorar a inteligibilidade e a qualidade do sinal de voz degradadas por ruído e reverberação. Também foi mostrado que as máscaras acústicas não-ideais BRM e BAM, mesmo não tenho sido desenvolvidas para ambientes ruidosos e reverberados simultaneamente, obtiveram exito em melhorar a inteligibilidade e a qualidade em algumas situações. Além disso, foi realizado um estudo da estacionariedade dos sinais de voz nestes cenários através do INS, que mostrou como a presença de ruído e reverberação aumenta a estacionariedade do sinal de voz. Estes resultados sugerem que o INS possa ser utilizado em conjunto com as máscaras acústicas como critério de seleção para determinar a predominância da interferência em quadros tempo-frequência.

REFERÊNCIAS

[1] D. Wang, U. Kjems, M. S. Pedersen, J. B. 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, no. 4, pp. 2336–2347, 2009.

[2] Y. Zaltz, Y. Bugannim, D. Zechoval, L. Kishon-Rabin, and R. Perez, "Listening in noise remains a significant challenge for cochlear implant users: Evidence from early deafened and those with progressive hearing loss compared to peers with normal hearing," *Journal of Clinical Medicine*, vol. 9, no. 5, 2020.

[3] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, pp. 117–128, 01 2000.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*. USA: CRC Press, Inc., 2nd ed., 2013.

[5] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7092–7096, 2013.

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

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] E. W. Healy and J. L. Vasko, "An ideal quantized mask to increase intelligibility and quality of speech in noise," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1392–1405, 2018.

[9] O. Hazrati, J. Lee, and P. C. Loizou, "Binary mask estimation for improved speech intelligibility in reverberant environments," in *INTERSPEECH*, 2012.

[10] 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.

[11] 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.

[12] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3988–3997, 2006.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4214–4217, 2010.

[14] C. Colomes, C. Schmidmer, T. Thiede, and W. C. Treurniet, "perceptual quality assessment for digital audio: PEAQ-the new ITU standard for objective measurement of the perceived audio quality," *journal of the audio engineering society*, september 1999.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, *Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs*, vol. 2, pp. 749–752. 2001.

[16] P. I. M. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," pp. 58–69, 1972.

[17] R. D. Patterson and B. C. J. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," *Frequency selectivity in hearing*, pp. 123–177, 1986.

[18] D. Pastor and F.-X. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences," *IEEE Transactions on Signal Processing*, vol. 60, pp. 1545–1555, 2012.

[19] K. Rhebergen and N. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 117, pp. 2181–92, 05 2005.

[20] C. Pavlovic, "SII—speech intelligibility index standard: Ansi s3.5 1997," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1906–1906, 2018.

[21] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, 2013.

[22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 01 1993.

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

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