Métodos de Mitigação de Efeitos Reverberantes e Ruidosos com Ganho de Inteligibilidade e Qualidade

A. Queiroz e R. Coelho

***Resumo*— Este artigo apresenta uma análise de métodos de aprimoramento da inteligibilidade e qualidade de sinais de voz em cenários reverberantes e ruidosos. Quatro métodos são considerados na avaliação para três diferentes condições de reverberação e quatro ruídos com diferentes graus de não-estacionariedade. Duas medidas objetivas de predição de inteligibilidade e duas de qualidade são aplicadas para avaliação objetiva. Os resultados mostram que a solução GTF_{F0} (*F0-based Gammatone Filtering*) alcança maior aprimoramento tanto de inteligibilidade quanto na qualidade acústica, quando comparada com outros métodos competitivos.**

***Palavras-Chave*— Inteligibilidade e qualidade acústica, não-estacionariedade, sinais reverberantes e ruidosos.**

***Abstract*— This work presents an analysis of methods to improve intelligibility and quality of speech signals under noisy reverberant scenarios. Four methods are considered in the evaluation for three different reverberation rooms and four noises with different degrees of nonstationarity. Two objective intelligibility measures and two quality measures are applied for objective assessment. Results show that the GTF_{F0} (*F0-based Gammatone Filtering*) solution leads to the highest improvement in terms of intelligibility and quality.**

***Keywords*— Acoustic intelligibility and quality, noisy reverberant signals, nonstationarity.**

I. Introdução

Reverberação é um efeito comumente presente em ambientes urbanos, abertos e fechados, tais como parques, auditórios e escritórios. Este efeito pode alterar as características da voz, impactando a qualidade e inteligibilidade [1][2]. Além disso, ruídos acústicos de diferentes graus de não-estacionariedade estão presentes em ambientes urbanos e podem corromper os sinais de voz. Quando presentes em um mesmo ambiente, tais efeitos compõem cenários reverberantes e ruidosos, os quais tornam-se um desafio para o aprimoramento da qualidade e inteligibilidade do sinal de voz.

Na literatura, métodos de realce [3][4][5] foram desenvolvidos para atuar em ambientes com ruídos não-estacionários. Estas soluções buscam estimar as estatísticas dos ruídos e, em seguida, aprimorar o sinal corrompido, obtendo resultados interessantes de qualidade e preservando a inteligibilidade. Em contrapartida, máscaras acústicas [6][7][8] são estudadas com enfoque no aprimoramento da inteligibilidade. Estas estratégias buscam emular o efeito *Cocktail Party* [9], o qual consiste na capacidade humana de segregar um sinal de interesse dentre múltiplas fontes. Além das técnicas de realce e máscaras acústicas, outras soluções foram propostas recentemente com o intuito de prover tanto inteligibilidade quanto qualidade aos sinais de voz ruidosos. Tais métodos, como o SSFV (*Smoothed Shifting of Formants for Voiced segments*) [10], $\text{APES}_{\text{HARM}}$ (*HARMonic-based Amplitude and Phase EStimation*) [11] e GTF_{F0} (*F0-based Gammatone Filtering*) [12], atuam nos componentes harmônicos do sinal, como a frequência fundamental (F0) e formantes. Recentemente, outras soluções [13][14] têm sido propostas levando em consideração cenários mais desafiadores, compostos por sinais reverberantes e ruidosos.

O presente artigo propõe a análise dos métodos SSFV, $\text{APES}_{\text{HARM}}$ e GTF_{F0} para sinais reverberantes e ruidosos. Estas soluções foram inicialmente investigadas em [12] com enfoque no aprimoramento da inteligibilidade de sinais somente para cenários ruidosos. Neste estudo, os métodos são avaliados também considerando os efeitos da reverberação. Além disso, os ganhos dos filtros *Gammatone* do GTF_{F0} são ajustados como contribuição para prover incrementos na qualidade, sem prejudicar a inteligibilidade aprimorada dos sinais. Para melhor avaliar o ganho de qualidade obtido, a solução de realce SCOE (*Single-Channel Online Enhancement*) [13] é também adotada como método comparativo.

Extensivos experimentos são realizados para avaliar objetivamente os métodos comparativos em termos de aprimoramento de inteligibilidade e qualidade da voz. Os cenários são compostos pela resposta ao impulso (*Room Impulse Response* - RIR) de três salas reais de reverberação: *Meeting* e *Stairway* da base AIR [15], e LASP2 da base LASP_RIR[1]. Ademais, foram selecionados quatro ruídos (Balbúrdia, Cafeteria, SSN - *Speech Shaped Noise* e Helicóptero) com valores de razão sinal-ruído (*Signal-to-Noise Ratio* - SNR) de -5 dB, 0 dB e 5 dB. Quatro medidas objetivas são consideradas na avaliação: ESTOI (*Extended Short-Time Objective Intelligibility*) [16] e ASII_{ST} (*short-time Approximated Speech Intelligibility Index*) [17] para predição de inteligibilidade, e PEAQ (*Perceptual Evaluation of Audio Quality*) [18] e PESQ (*Perceptual Evaluation of Speech Quality*) [19] para qualidade dos sinais. Os resultados dos experimentos demonstram que a solução GTF_{F0} obteve maior aprimoramento nas métricas avaliadas, superando tanto em inteligibilidade quanto em qualidade os demais métodos competitivos.

O restante do artigo está organizado da seguinte forma: A Seção II demonstra os efeitos causados pela reverberação e ruídos na voz, por meio dos seus índices de não-estacionariedade. A Seção III descreve os métodos adotados para aprimoramento da qualidade e inteligibilidade acústica. Na Seção IV, as técnicas são avaliadas em diferentes cenários reverberantes e ruidosos, por meio de medidas objetivas de inteligibilidade e qualidade. Enfim, 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 (308155/2019-0) e pela FAPERJ (203075/2016). E-mails: {anderson.queiroz,coelho}@ime.eb.br.

[1]Disponível em lasp.ime.eb.br

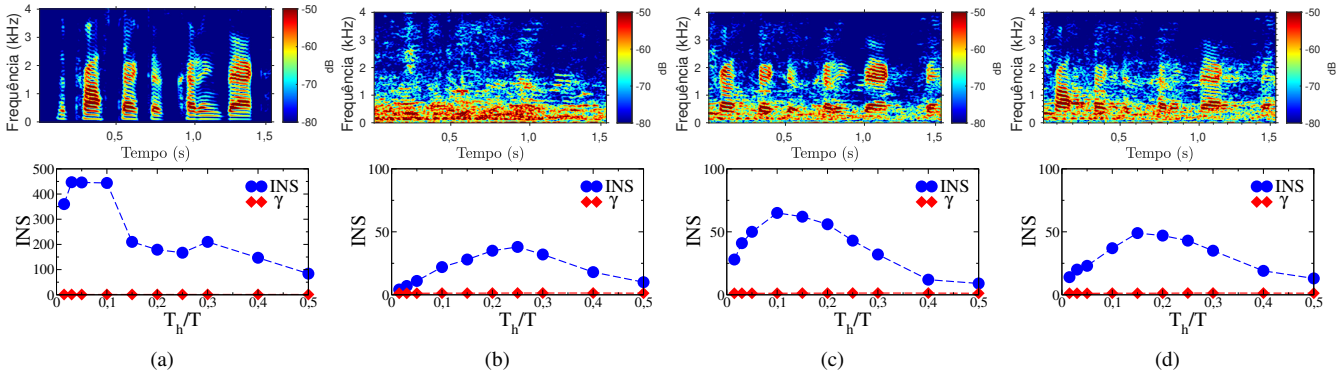

Fig. 1. Espectrogramas e respectivos valores de INS para: (a) segmento de voz limpo, (b) ruído Balbúrdia, e sinais de voz reverberantes e ruidosos com ruído Balbúrdia (SNR=0 dB) para as salas (c) *Meeting* (RT_{60} = 0,36 s) e (d) *Stairway* (RT_{60} = 1,05 s).

II. Efeitos da Reverberação e Ruídos na Voz

Um sinal reverberado pode ser obtido pela convolução do sinal de voz com a resposta ao impulso $h(t)$ de uma sala. A RIR é caracterizada pelo tempo de reverberação RT_{60} e pela razão entre os sinais direto e reverberado (*Direct-to-Reverberant Ratio* - DRR). Deste modo, um sinal reverberante e ruidoso pode ser descrito por $x(t) = s(t) * h(t) + w(t)$, onde $s(t)$ é a voz, e $w(t)$ o ruído acústico.

O Índice de Não-Estacionariedade (*Index of NonStationarity* - INS) [20] é adotado na análise dos sinais de voz reverberantes e ruidosos. O INS é obtido a partir da comparação do sinal de voz com referenciais estacionários chamados *surrogates*. Para isto, a escala de observação T_h/T é definida como a razão entre o tamanho da janela utilizada na análise espectral (T_h), e a duração total do sinal (T). Em [20], um limiar é definido para cada valor da janela T_h, considerando uma precisão de 95%. Desta forma, a avaliação da hipótese de estacionariedade é realizada comparando o limiar com o valor de INS, ou seja

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

A Figura 1 mostra os espectrogramas e valores de INS para quatro cenários: sinal limpo (a), ruído Balbúrdia (b), e reverberantes e ruidosos com as salas (c) *Meeting* (RT_{60} = 0,36 s) e (d) *Stairway* (RT_{60} = 1,05 s), com SNR 0 dB. Note que o sinal limpo possui suas componentes de frequência bem definidas, ao contrário das locuções reverberantes e ruidosas que apresentam distorções em todo o espectro, principalmente nas frequências abaixo de 1 kHz. Segundo os resultados de INS, os sinais nas quatro condições apresentam não-estacionariedade em todas as escalas temporais. Note na Figura 1(a) que o INS evidencia a natureza altamente não-estacionária do sinal de voz limpo, cujo $\text{INS}_{\text{máx}}$ atingiu 450. Já o ruído Balbúrdia (Figura 1(b)), obteve o valor de INS máximo de 38 sendo classificado como não-estacionário. O sinal de voz teve sua propriedade não-estacionária atenuada quando afetados por ambientes reverberantes e ruidosos, atingindo $\text{INS}_{\text{máx}}$ de 65 e 48 para as salas *Meeting* e *Stairway*, respectivamente. Estes resultados demonstram o impacto dos efeitos da reverberação e do ruído no sinal de voz, que consequentemente podem reduzir a qualidade e inteligibilidade acústica [1][5].

III. Soluções para Aprimoramento da Inteligibilidade e Qualidade Acústica

Esta Seção descreve brevemente os métodos de aprimoramento de inteligibilidade e qualidade SSFV, $\text{APES}_{\text{HARM}}$, SCOE e GTF_{F0} investigados neste trabalho.

1) **SSFV:** Este método [10] considera o mascaramento acústico em ambientes ruidosos como o principal responsável pela redução da inteligibilidade de sinais de voz. Nestes cenários, seres humanos naturalmente alteram a produção da voz, tornando-a audível por meio da utilização de alguns artifícios, os quais são representados pelo termo conhecido como Efeito Lombard [23] [24]. A principal ideia desta solução consiste em aprimorar a inteligibilidade do sinal, deslocando as frequências centrais das formantes (*Formant Shifting*) de acordo com uma função proposta em [25], distanciando tais frequências da região de atuação espectral do ruído.

2) **$\text{APES}_{\text{HARM}}$:** Esta técnica [11] adota o filtro APES [26], de modo que a atuação ocorra nos sinais harmônicos ($\text{APES}_{\text{HARM}}$). Em cada segmento sonoro (*voiced*), a F0 é estimada e os harmônicos são definidos como seus múltiplos inteiros ($f_c = F0, 2F0, \ldots, LF0$). Na sequência, o banco de filtros APES é implementado, com as frequências centrais definidas por f_c. A reconstrução do sinal é realizada utilizando os segmentos sonoros filtrados, e os segmentos surdos (*unvoiced*) originais do sinal de voz. O sinal processado apresenta um aumento da SNR quando comparado com o sinal original, visto que o filtro APES atenua o efeito de algumas regiões espectrais do ruído.

3) **SCOE:** A solução SCOE [13] utiliza uma abordagem de filtragem Bayesiana para o realce de sinais de voz em ambientes reverberantes e ruidosos, obtendo uma estimação conjunta de parâmetros acústicos e energia da voz, ruído e reverberação na escala Mel. O método considera o modelo HMM (*Hidden Markov Model*) para a produção da voz e um modelo autorregressivo para a energia de reverberação. A partir dessas estimativas, um ganho espectral é definido e o aprimoramento espectral é aplicado no domínio da STFT (*Short Time Fourier Transform*).

4) **GTF_{F0}:** A principal ideia do método proposto em [12] consiste em filtrar o sinal com filtros *Gammatone* [27], utilizando as estimativas da F0 do sinal e seus múltiplos harmônicos. Para quadros sonoros, cada filtro é implementado[2]

[2]Disponível em http://staffwww.dcs.shef.ac.uk/people/N.Ma/.

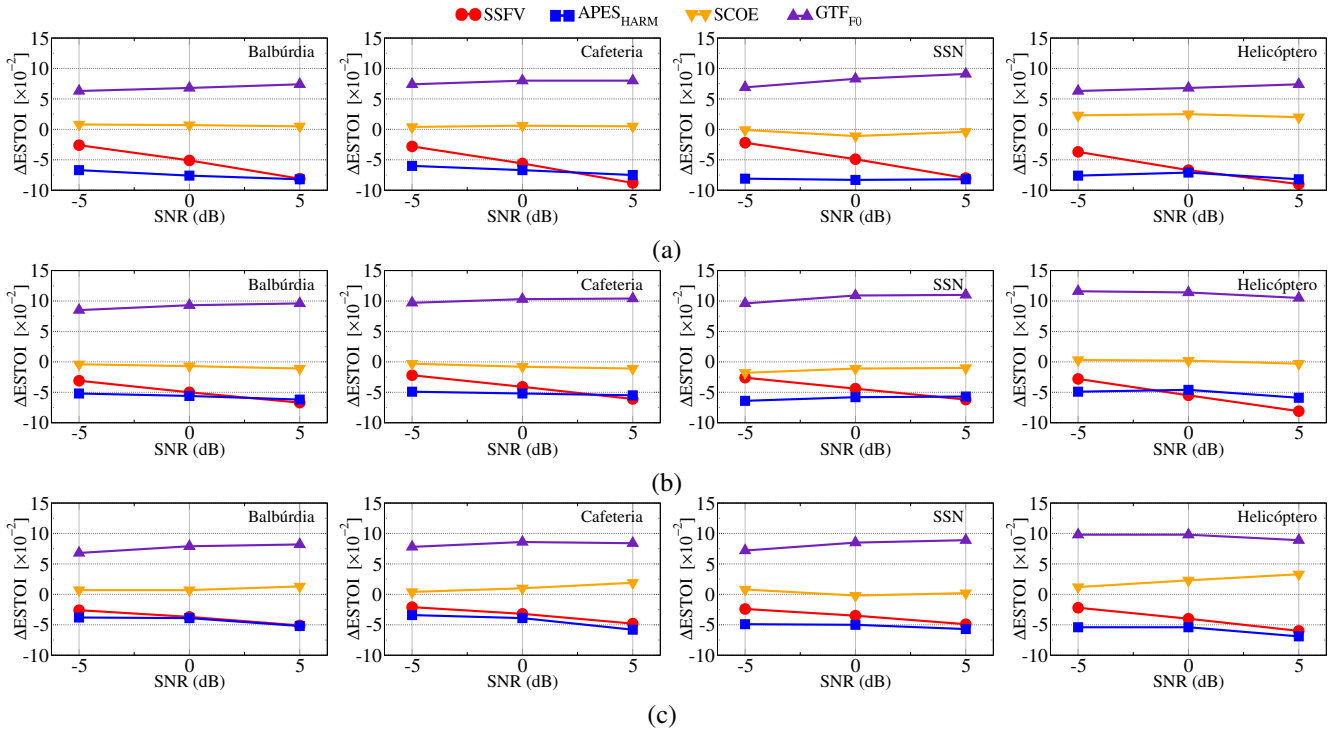

Fig. 2. Aprimoramento da inteligibilidade ΔESTOI [$\times 10^{-2}$] dos cenários reverberantes e ruidosos para as salas *Meeting* (a), LASP2 (b) e *Stairway* (c).

no domínio do tempo, considerando as frequências centrais dadas por $f_c = kF0$, $k = 1, \ldots, L$. Após a filtragem, a amplitude das amostras na saída de cada filtro é amplificada para salientar a presença dos L harmônicos da frequência fundamental. Assim como no APES$_{\text{HARM}}$, o sinal é reconstruído agrupando os segmentos processados com os demais segmentos do sinal original.

IV. Experimentos, Resultados e Discussão

Nesta Seção, os métodos SSFV, APES$_{\text{HARM}}$, SCOE e GTF$_{\text{F0}}$ são avaliados em termos de inteligibilidade e qualidade, considerando vários cenários reverberantes e ruidosos. Um subconjunto [28] da base TIMIT [29] foi selecionado para os experimentos. Este subconjunto é composto por 128 sinais de 16 locutores (8 masculinos e 8 femininos), com taxa de amostragem de 16 kHz. Três salas foram selecionadas para representar os efeitos da reverberação em ambientes urbanos reais: a sala LASP2 da base LASP_RIR, e as salas *Meeting* e *Stairway* da base AIR. As salas *Meeting*, LASP2 e *Stairway* apresentam valores de RT_{60} e DRR de {0,36;0,79;1,05} e {1,22;-4,37;-5,28}, respectivamente. Finalmente, os sinais de voz reverberados são corrompidos pelos ruídos Balbúrdia da base RSG-10 [31]; Cafeteria e Helicóptero da base Freesound.org[3]; e SSN da base DEMAND [30]. Os ruídos Balbúrdia e Cafeteria apresentam INS$_{\text{máx}}$ de 38 e 23, ou seja, são não-estacionários. Por outro lado, Helicóptero e SSN são estacionários, com INS$_{\text{máx}}$ de 2 e 1, respectivamente.

Os métodos APES$_{\text{HARM}}$ e GTF$_{\text{F0}}$ foram implementados considerando a técnica de estimação de F0 HHT-Amp [21]. Estudos mostram que esta solução supera outros estimadores em diversos cenários ruidosos [21] [22] e reverberantes e ruidosos [32]. Assim como em [12] a solução GTF$_{\text{F0}}$ foi aplicada para quadros de 32 ms, considerando filtros de ordem n = 4 e largura de banda $b = 0,25F0$. No entanto, neste estudo os ganhos lineares G_k foram ajustados para os nove primeiros múltiplos da F0, ou seja, $G_k = \{5; 5; 4; 3; 2, 5; 2; 1, 5; 1, 5; 1, 5\}$ para $k \in \{1; \ldots; 9\}$. Este novo conjunto de valores torna o GTF$_{\text{F0}}$ robusto para atuação em cenários reverberantes e ruidosos. Além de prover inteligibilidade acústica, este ajuste nos ganhos também aprimora a qualidade dos sinais.

[3]Disponível em www.freesound.org.

TABELA I

ESTOI dos sinais Não Processados (NP) reverberados com as três salas, e corrompidos por quatro ruídos.

	Meeting			LASP2			*Stairway*		
SNR (dB)	-5	0	5	-5	0	5	-5	0	5
Balbúrdia	0,25	0,37	0,49	0,30	0,41	0,50	0,22	0,33	0,44
Cafeteria	0,27	0,39	0,53	0,33	0,43	0,52	0,24	0,35	0,46
SSN	0,24	0,36	0,49	0,30	0,40	0,49	0,22	0,33	0,44
Helicóptero	0,37	0,49	0,60	0,40	0,48	0,55	0,34	0,44	0,53

As medidas ESTOI [16] e ASII$_{\text{ST}}$ [17] foram adotadas para avaliação objetiva de inteligibilidade. A primeira realiza um cálculo da correlação espectral entre os coeficientes de diferentes bandas de frequência. A medida ASII$_{\text{ST}}$ foi desenvolvida para ser capaz de lidar com o comportamento não-estacionário dos ruídos acústicos, e também com os efeitos da reverberação. Ambas as medidas apresentam valores compreendidos entre 0 e 1, onde a maior inteligibilidade é indicada por resultados próximos de 1. A Tabela I apresenta os resultados médios de ESTOI para os sinais reverberantes e ruidosos Não Processados (NP). A sala *Stairway* apresenta o cenário mais desafiador com o maior valor de RT_{60} e menor DRR. Note que o menor valor de ESTOI é obtido nesta sala (ESTOI = 0,22) para os ruídos Balbúrdia e SSN com SNR = -5 dB.

A Figura 2 mostra o aprimoramento médio da inteligibilidade ESTOI (ΔESTOI) para as três salas e quatro ruídos.

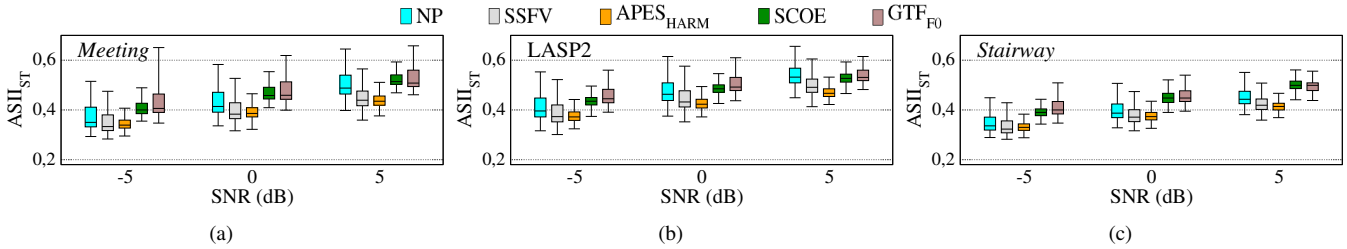

Fig. 3. Resultados de inteligibilidade $ASII_{ST}$ dos cenários reverberantes e ruidosos considerando os ruídos Balbúrdia, Cafeteria, SSN e Helicóptero.

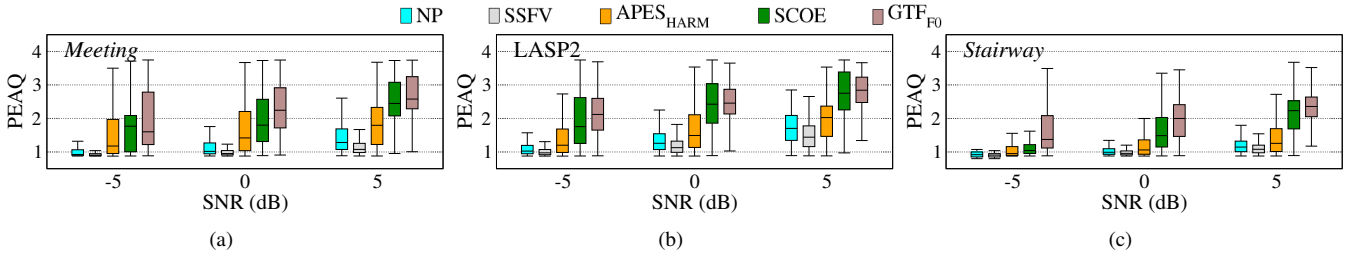

Fig. 4. Resultados de qualidade PEAQ para os cenários reverberantes e ruidosos considerando os ruídos Balbúrdia, Cafeteria, SSN e Helicóptero.

TABELA II

RESULTADOS DE PESQ PARA AS SALAS *Meeting*, LASP2 E *Stairway* COM OS RUÍDOS BALBÚRDIA, CAFETERIA, SSN E HELICÓPTERO.

		Balbúrdia				Cafeteria				SSN				Helicóptero				Média
	SNR (dB)	-5	0	5	Méd.	-5	0	5	Méd.	-5	0	5	Méd.	-5	0	5	Méd.	Total
Meeting	NP	1,25	1,56	1,87	1,56	1,39	1,70	2,01	1,70	1,19	1,50	1,83	1,51	1,50	1,81	2,11	1,81	1,64
	SSFV	1,31	1,58	1,82	1,57	1,40	1,65	1,90	1,65	1,29	1,53	1,78	1,53	1,47	1,74	1,99	1,73	1,62
	$APES_{HARM}$	1,05	1,31	1,63	1,33	1,18	1,48	1,80	1,48	0,99	1,30	1,63	1,31	1,31	1,65	1,94	1,63	1,44
	SCOE	1,24	1,61	1,96	1,60	1,35	1,72	2,08	1,72	1,27	1,59	1,98	1,61	1,56	1,92	2,23	1,90	1,71
	GTF_{F0}	**1,35**	**1,71**	**2,08**	**1,71**	**1,54**	**1,92**	**2,26**	**1,90**	**1,29**	**1,66**	**2,06**	**1,67**	**1,70**	**2,07**	**2,39**	**2,05**	**1,84**
LASP2	NP	1,35	1,66	1,96	1,66	1,44	1,77	2,06	1,76	1,30	1,61	1,93	1,61	1,58	1,89	2,16	1,88	1,73
	SSFV	1,42	1,70	1,93	1,68	1,45	1,76	1,99	1,73	1,37	1,64	1,90	1,64	1,59	1,86	2,08	1,84	1,72
	$APES_{HARM}$	1,25	1,56	1,84	1,55	1,34	1,67	1,94	1,65	1,19	1,51	1,83	1,51	1,49	1,80	2,00	1,76	1,62
	SCOE	1,29	1,59	1,89	1,59	1,33	1,63	1,95	1,64	1,21	1,58	1,89	1,56	1,55	1,87	2,11	1,84	1,66
	GTF_{F0}	**1,46**	**1,84**	**2,19**	**1,83**	**1,64**	**2,01**	**2,32**	**1,99**	**1,41**	**1,81**	**2,16**	**1,79**	**1,81**	**2,15**	**2,42**	**2,12**	**1,93**
Stairway	NP	1,26	1,56	1,83	1,55	1,37	1,66	1,91	1,64	1,22	1,53	1,82	1,52	1,51	1,81	2,03	1,78	1,62
	SSFV	1,34	1,61	1,84	1,59	1,40	1,66	1,89	1,65	1,31	1,57	1,82	1,57	1,52	1,79	1,98	1,76	1,64
	$APES_{HARM}$	1,13	1,45	1,72	1,43	1,27	1,55	1,78	1,54	1,15	1,45	1,72	1,44	1,42	1,71	1,89	1,67	1,52
	SCOE	1,27	1,62	1,96	1,62	1,42	1,75	2,06	1,74	**1,35**	1,64	1,99	1,66	1,57	1,93	2,20	1,90	1,73
	GTF_{F0}	**1,37**	**1,73**	**2,03**	**1,71**	**1,54**	**1,87**	**2,12**	**1,84**	1,34	**1,72**	**2,02**	**1,69**	**1,72**	**2,03**	**2,24**	**2,00**	**1,81**

Observe que o método GTF_{F0} apresenta os maiores incrementos de inteligibilidade em todos os cenários. O maior valor de ΔESTOI ($11{,}6\times10^{-2}$) foi alcançado na sala LASP2 pelo ruído Helicóptero com SNR = -5 dB. Para a sala *Stairway* com o ruído não-estacionário Balbúrdia com SNR = 0 dB, GTF_{F0} atingiu aprimoramento de $7{,}9\times10^{-2}$, superando os valores de $0{,}7\times10^{-2}$, $-3{,}7\times10^{-2}$ e $-3{,}9\times10^{-2}$ para os métodos SCOE, SSFV e $APES_{HARM}$, respectivamente.

A Figura 3 ilustra os resultados das avaliações objetivas de inteligibilidade com a medida $ASII_{ST}$ para os cenários provenientes das três salas de reverberação. Cada diagrama de caixa representa a distribuição dos resultados para os quatro ruídos (Balbúrdia, Cafeteria, SSN e Helicóptero). Os resultados de $ASII_{ST}$ mostram que os métodos GTF_{F0} e SCOE alcançam os maiores valores médios de inteligibilidade para a maioria das composições reverberantes e ruidosas. Este caso está evidenciado para a sala *Meeting* em -5 dB, na qual ambas as técnicas atingiram $ASII_{ST}$ = 0,40. Para o método GTF_{F0}, nota-se valores médios de inteligibilidade de 0,49, 0,52 e 0,46 para as salas *Meeting*, LASP2 e *Stairway*, seguido de 0,47, 0,48 e 0,45 para a estratégia SCOE. Os métodos SSFV e $APES_{HARM}$ não superaram os demais métodos em nenhum dos cenários. Isto se deve ao fato de que estes métodos não levam em consideração o impacto da reverberação nos componentes harmônicos da voz.

A avaliação do aprimoramento da qualidade obtido pelos métodos comparativos é realizada a partir das medidas PEAQ [18] e PESQ [19]. Ambas as medidas estão apresentadas na escala MOS (*Mean Opinion Score*), cujos valores próximos de 1,0 indicam baixa qualidade acústica e acima de 4,0 o sinal apresenta pouca ou nenhuma distorção (alta qualidade). A Figura 4 mostra os resultados de PEAQ onde novamente cada diagrama de caixa denota os resultados para os quatro ruídos. Observe que a solução GTF_{F0} supera em termos de valores médios a estratégia de realce SCOE na maioria dos cenários, exceto para a sala *Meeting* com SNR = -5 dB. Assim como GTF_{F0} e SCOE, a solução $APES_{HARM}$ também apresentou incrementos nos valores de PEAQ. Por exemplo, note que para a sala LASP2, GTF_{F0} apresenta PEAQ médio de 2,12, 2,46 e 2,84 para -5 dB, 0 dB e 5 dB, seguido de 1,76, 2,43 e 2,75 para SCOE e 1,20, 1,49 e 2,03 para a solução $APES_{HARM}$.

A Tabela II apresenta as predições de PESQ obtidas para os cenários experimentais. Note que o método GTF_{F0} atinge os maiores resultados em todos os valores de SNR. Para o ruído

Balbúrdia na sala *Meeting*, este método apresenta média de PESQ de 1,71, seguido de 1,60 e 1,57 para SCOE e SSFV, respectivamente. Para a sala LASP2, o melhor desempenho é apresentado por GTF_{F0}, acompanhado de SSFV e SCOE. Estas soluções atingem resultados médios de 1,83, 1,68 e 1,59 para o ruído Balbúrdia e 1,79, 1,64 e 1,56 para SSN. Para a sala *Stairway*, considerado o ambiente mais desafiador, a solução GTF_{F0} obtém os maiores resultados médios de PESQ de 2,00 e 1,84 para os ruídos Helicóptero e Cafeteria, respectivamente. Nestes cenários, SCOE apresenta valores de 1,90 e 1,74 seguido do SSFV com 1,76 e 1,65. Estes resultados reforçam a capacidade do método GTF_{F0} de prover também qualidade acústica aos sinais de voz em ambientes reverberantes e ruidosos.

V. Conclusão

Este artigo apresentou um estudo comparativo de quatro métodos de aprimoramento da inteligibilidade e qualidade da voz para aplicações em ambientes reverberantes e ruidosos. Extensivos experimentos foram conduzidos utilizando três salas de reverberação e quatro ruídos de distintos graus de não-estacionariedade. Os resultados mostraram que o método GTF_{F0} apresentou os melhores resultados de aprimoramento da inteligibilidade dos sinais para diferentes cenários. Além disso, os resultados objetivos obtidos pelo GTF_{F0} também demonstraram expressivo aprimoramento na qualidade dos sinais. Neste contexto, o resultado médio de PESQ foi aprimorado em 11,8%, seguido de 2,2% pela solução de realce SCOE. Portanto, o método GTF_{F0} foi capaz de aprimorar tanto a inteligibilidade quanto a qualidade dos sinais, superando os demais métodos comparativos em cenários desafiadores compostos por reverberação e ruído.

Referências

[1] R. J. Bolt, e A. D MacDonald, "Theory of speech masking by reverberation," *The Journal of the Acoustical Society of America*, v. 21, no. 6, pp. 577–580, 1949.

[2] A. Nabelek, "Communication in noisy and reverberant environments," *Acoustical factors affecting hearing aid performance*, pp. 15–28, 1993.

[3] T. Gerkmann, e R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, v. 20, no. 4, pp. 1383–1393, 2012.

[4] 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*, v. 22, no. 5, pp. 899–911, 2014.

[5] R. Tavares, e R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Processing Letters*, v. 23, no. 1, pp. 6–10, 2016.

[6] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, Boston, MA: Springer, pp. 181–197, 2005.

[7] L. Zão, D. Cavalcante, e R. Coelho, "Time-frequency feature and AMS-GMM mask for acoustic emotion classification," *IEEE Signal Processing Letters*, v. 21, no. 5, pp. 620–624, 2014.

[8] F. Farias, e R. Coelho, "Blind Adaptive Mask to Improve Intelligibility of Non-Stationary Noisy Speech," *IEEE Signal Processing Letters*, v. 28, pp. 1170–1174, 2021.

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

[10] K. Nathwani, G. Richard, B. David, P. Prablanc e V. Roussarie, "Speech intelligibility improvement in car noise environment by voice transformation," *Speech Communication*, v. 91, pp. 17–27, Maio 2017.

[11] S. Norholm, J. Jensen e M. Christensen, "Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech," *IEEE Transactions on Audio, Speech and Language Processing*, v. 24, no. 4, pp. 645–658, Abril 2016.

[12] 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.

[13] C. S. J. Doire *et al*, "Single-Channel Online Enhancement of Speech Corrupted by Reverberation and Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 25, no. 3, pp. 572–587, 2017.

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

[15] M. Jeub, M. Schaefer, e 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, Jul. 2009.

[16] J. Jensen e C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, v. 24, no. 11, pp. 2009–2022, 2016.

[17] R. C. Hendriks, J. B. Crespo, J. Jensen e C. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, v. 23, no. 5, pp. 851–862, 2015.

[18] C. Colomes, C. Schmidmer, T. Thiede e 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*, 1999.

[19] A. Rix, J. Beerends, M. Hollier e A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, v. 2, pp. 749–752, 2001.

[20] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, e J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, v. 58, no. 7, pp. 3459-3470, Jul. 2010.

[21] L. Zão e R. Coelho, "On the Estimation of Fundamental Frequency From Nonstationary Noisy Speech Signals Based on the Hilbert–Huang Transform," *IEEE Signal Processing Letters*, v. 25, no. 2, pp. 248–252, Fevereiro. 2018.

[22] A. Queiroz e R. Coelho, "Estimação de Frequência Fundamental de Sinais Acústicos Ruidosos com Aprendizado de Máquina," *XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT*, 2021.

[23] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, v. 37, pp. 101–119, 1911.

[24] J. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, v. 93, no. 1, pp. 510–524, 1993.

[25] K. Nathwani, M. Daniel, G. Richard, B. David e V. Roussarie, "Formant Shifting for Speech intelligibility improvement in car noise environment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5375–5379, 2016.

[26] P. Stoica, H. Li e J. Li, "A New Derivation of the APES Filter," *IEEE Signal Processing Letters*, v. 6, no. 8, pp. 205–206, Agosto 1999.

[27] J. Holdsworth, I. Nimmo-Smith, R. Patterson and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, v. 1, pp. 1–5, 1988.

[28] S. Gonzalez, "Pitch of the core timit database set," *Disponível em: <http://www.ee.ic.ac.uk/hp/staff/dmb/data/TIMITfxv.zip>*, 2014.

[29] S. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, e 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*, v. 24, 1993.

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

[31] 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.

[32] 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.