Supressão de Eco Acústico por Máscara Tempo-Frequência e Redes Neurais Recorrentes

Erik S. Milesi e Bruno C. Bispo

***Resumo*— Este artigo avalia a combinação de redes neurais recorrentes (RNNs) e máscaras tempo-frequência na supressão de eco acústico. Quatro arquiteturas de RNNs são propostas para estimar a máscara de razão ideal, a qual é aplicada para suprimir o eco em situações de conversação cruzada. Os resultados indicam que, nos trechos de fala única, modelos causais e de baixo custo computacional conseguem atenuar o eco em 20, 34 e 49 dB para níveis de razão sinal-eco de 0, 3,5 e 5 dB, respectivamente, enquanto modelos não-causais e de alto custo computacional alcançam resultados pelo menos 15 dB superiores. No entanto, o melhor desempenho dos modelos de alto custo computacional não se reflete significativamente na qualidade e na inteligibilidade dos trechos de conversação cruzada.**

***Palavras-Chave*— Eco acústico, supressão de eco, conversação cruzada, máscara tempo-frequência, redes neurais.**

***Abstract*— This paper evaluates the combination of recurrent neural networks (RNNs) and time-frequency masks in acoustic echo suppression. Four RNN architectures are proposed to estimate the ideal ratio mask, which is applied to suppress echo in double-talk situations. The results indicate that, in the single-talk segments, low-computational cost models may attenuate the echo by 20, 34 and 49 dB for signal-to-echo ratios of 0, 3.5 and 5 dB, respectively, while high computational cost models achieve results at least 15 dB higher. However, the better performance of high-computational cost models is not significantly reflected in the quality and intelligibility of double-talk segments.**

***Keywords*— Acoustic echo, echo supression, double talk, time-frequency mask, neural networks.**

I. Introdução

Nos sistemas de comunicação de mãos livres, o acoplamento acústico entre alto-falante e microfone numa extremidade da conexão, chamada de próxima, inevitavelmente faz o sinal reproduzido pelo alto-falante ser capturado e transmitido de volta para a outra extremidade, chamada de distante. Assim um locutor na extremidade distante recebe sua própria fala. Como o atraso causado pela transmissão é geralmente de centenas de milissegundos, a réplica da fala do locutor distante é distinguível da sua fala original e soa como eco. O eco acústico é transmitido junto aos sinais sonoros gerados na extremidade próxima, incluindo sinais de fala se houver comunicação simultânea, também chamada de conversação cruzada.

As soluções do estado-da-arte para o problema em questão são os métodos de cancelamento de eco acústico (AEC, do inglês *acoustic echo cancellation*). Esses métodos visam estimar o eco acústico e subtraí-lo do sinal do microfone. A estimativa do eco é obtida ao filtrar o sinal do alto-falante próximo com um modelo do caminho do eco acústico. Este modelo é calculado usando um filtro adaptativo que foi projetado para estimar e rastrear esse caminho por meio de um algoritmo recursivo. Esses algoritmos são chamados algoritmos de filtragem adaptativa.

No entanto, se forem utilizados os algoritmos tradicionais de filtragem adaptativa, que aproximam a solução de Wiener através do gradiente ou mínimos quadrados, o filtro adaptativo não deve ser atualizado quando há conversação cruzada porque o sinal do locutor próximo age como ruído para o filtro ao impedir que o sinal de eco compensado seja nulo. De fato, tanto o ruído ambiente quanto o sinal do locutor próximo atuam como ruído para o filtro, mas o último é muito mais prejudicial devido à sua maior intensidade. Como consequência, o sinal do locutor próximo atrapalha a atualização do filtro e pode causar sua divergência. Um detector de conversação cruzada é normalmente empregado para indicar essa situação.

Muito trabalho tem sido dedicado ao desenvolvimento de filtros adaptativos para cancelamento, ou pelo menos supressão, do eco acústico e ao projeto de mecanismos de controle sofisticados para garantir robustez em condições adversas [1]. No entanto, estudos recentes demonstraram que a separação da fala como um problema de aprendizagem supervisionada tem se mostrado bastante promissora [2]. Nesse sentido, uma abordagem alternativa foi proposta recentemente em [3].

Essa abordagem consiste em um modelo de rede neural recorrente (RNN, do inglês *recurrent neural network*) para estimar a máscara tempo-frequência de razão ideal (IRM, do inglês *ideal ratio mask*), a qual é aplicada ao sinal do microfone para suprimir o eco acústico. Em condições de conversação cruzada, a abordagem superou o algoritmo NLMS (do inglês *normalized least mean squared*) quanto à atenuação do eco, mas não em termos de qualidade sonora [3]. Em cenários com conversação cruzada e ruído ambiente, a RNN superou o filtro adaptativo quanto à atenuação e qualidade. Apesar do desempenho promissor, a RNN proposta em [3] é não-causal, tem um alto custo computacional e alta latência, sendo assim inviável para aplicações em tempo real.

Para solucionar o desafio da aplicação em tempo real, este estudo analisou o desempenho de três modelos de RNN com baixo custo computacional e os comparou com o apresentado em [3], para analisar o compromisso entre desempenho, latência e custo computacional. Condições desafiadoras de conversação cruzada foram exploradas para avaliar a robustez e capacidade de generalização dos modelos.

Este artigo está organizado da seguinte forma: a Seção II descreve a metodologia proposta, apresentando brevemente a formulação do problema, a teoria das máscaras tempo-frequência, as técnicas de aprendizado de máquina empregadas, e as suas variáveis de entrada e de saída; a Seção III descreve a configuração das simulações computacionais realizadas; na Seção IV, os resultados obtidos são apresentados e

Erik S. Milesi, Centro de Guerra Acústica e Eletrônica da Marinha, Niterói-RJ, Brasil, e Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis-SC, Brasil. Bruno C. Bispo, Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal de Santa Catarina, Florianópolis-SC, Brasil. E-mails: milesi@marinha.mil.br, bruno.bispo@ufsc.br.

discutidos; por fim, a Seção V conclui o artigo.

II. Metodologia Proposta

Essa seção apresenta a formulação do problema de eco acústico e a metodologia proposta para sua supressão.

A. Formulação do Problema

O diagrama de blocos do problema é mostrado na Figura 1. Os sinais de fala do locutor distante e do locutor próximo são denotados por $x(n)$ e $s(n)$, respectivamente. O caminho do eco acústico é modelado pelo sistema linear e invariante no tempo com resposta ao impulso $h(n)$. O sinal de eco acústico é dado pela convolução $d(n) = x(n) * h(n)$ e o ruído aditivo é denotado por $v(n)$. O sinal de fala contaminado é definido como $y(n) = s(n) + d(n) + v(n)$. Os sinais $s(n)$, $d(n)$ e $v(n)$ são não-observáveis. Nesse trabalho é assumido $v(n) = 0$.

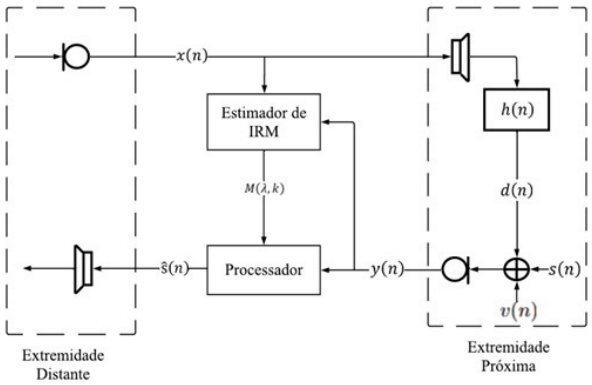

Fig. 1. Modelo de sinais nas comunicações bidirecionais de mãos livres.

A transformada de Fourier de tempo curto (STFT, do inglês *short-time Fourier transform*) de $y(n)$ é expressa como

$$Y(k,\lambda) = S(k,\lambda) + D(k,\lambda), \tag{1}$$

onde $S(k,\lambda)$ e $D(k,\lambda)$ são as STFTs de $s(n)$ e $d(n)$, λ é o índice da janela temporal e k é o bin de frequência.

B. Máscaras Tempo-Frequência

As máscaras tempo-frequência (MTFs) visam atenuar as faixas de frequência do sinal de fala contaminado dominadas pelo ruído. Isso é realizado ao multiplicar, a cada unidade tempo-frequência $\{k,\lambda\}$, $Y(k,\lambda)$ por uma máscara $M(k,\lambda)$, resultando numa estimativa de $S(k,\lambda)$ dada por [4]

$$\hat{S}(k,\lambda) = Y(k,\lambda)M(k,\lambda). \tag{2}$$

Apesar da razão sinal-ruído (SNR, do inglês *signal-to-noise ratio*) em cada unidade de tempo-frequência não ser alterada, a SNR global pode ser aumentada se ruído e fala ocuparem bandas de frequência diferentes ao longo do tempo. A estimativa $\hat{s}(n)$ no domínio do tempo do sinal do locutor próximo é reconstruída utilizando as STFTs inversas de $\hat{S}(k,\lambda)$ e uma estratégia de sobreposição-e-soma [5].

As máscaras podem ser definidas utilizando critérios objetivos ou heurísticas. Em geral, $0 \leq M(k,\lambda) \leq 1$ e $M(k,\lambda)$ é uma função da SNR a priori associada a λ-ésima janela e k-ésimo *bin* frequencial, a qual é definida como

$$\xi(k,\lambda) = \frac{S_s(k,\lambda)}{S_d(k,\lambda)}, \tag{3}$$

onde $S_s(k,\lambda) = \mathrm{E}\left\{|S(k,\lambda)|^2\right\}$ e $S_d(k,\lambda) = \mathrm{E}\left\{|D(k,\lambda)|^2\right\}$ são as densidades espectrais de potência de $s(n)$ e $d(n)$, respectivamente, e $\mathrm{E}\{\cdot\}$ é o operador valor esperado.

A máscara de Wiener (WM, do inglês *Wiener mask*) é a mais encontrada na literatura [4]. Ela é o filtro que minimiza o erro quadrático médio entre o sinal de fala e o sinal de fala contaminado, assumindo que fala e ruídos são independentes, ou pelo menos descorrelacionados, e com média zero.

Nos últimos anos, a máscara raiz de Wiener tem se destacado em diversas aplicações de fala. Ela é uma versão suavizada da WM com função de atenuação dada por [4]

$$M(k,\lambda) = \sqrt{\frac{\xi(k,\lambda)}{\xi(k,\lambda)+1}}. \tag{4}$$

Em implementações em tempo real, os valores esperados presentes em (4) são aproximados, levando à IRM [4].

C. Variáveis de Entrada

As variáveis de entrada dos modelos de aprendizado de máquina são provenientes de $x(n)$ e $y(n)$, os quais são amostrados a uma taxa de 16 kHz. Esses sinais são divididos em segmentos de 20 ms, com sobreposição de 10 ms, utilizando a janela de Hamming. Para cada segmento, a STFT com 320 pontos é calculada, resultando em 161 bins de frequência. Em seguida, a característica de magnitude logarítmica é obtida ao aplicar a operação logarítmica ao espectro de magnitude de cada segmento [6]. Dessa forma, para cada janela temporal, o resultado é um vetor de entrada com dimensão 161×2.

D. Variável Alvo

Em situações na qual se utilizou aprendizado profundo para estimar MTFs, a IRM proporcionou qualidade e inteligibilidade objetivas superior às demais [7]. Por esse motivo, como feito em [3], a variável alvo para os modelos de aprendizado de máquina será a máscara IRM dada por

$$M(k,\lambda) = \sqrt{\frac{\hat{\xi}(k,\lambda)}{\hat{\xi}(k,\lambda)+1}}. \tag{5}$$

onde

$$\hat{\xi}(k,\lambda) = \frac{|S(k,\lambda)|^2}{|D(k,\lambda)|^2} \tag{6}$$

é o valor instantâneo de $\xi(k,\lambda)$. Apesar de serem não-observáveis, os sinais $s(n)$ e $d(n)$ são utilizados no treinamento dos modelos de aprendizado de máquina.

E. Modelos de Aprendizado de Máquina

Os modelos de aprendizagem de máquina utilizados para estimar a IRM foram a memória longa de curto prazo (LSTM, do inglês *long short-time memory*), sua versão bidirecional (BLSTM, do inglês *bidirectional* LSTM) e rede neural convolucional recorrente (CRNN, do inglês *convolutional* RNN).

1) LSTM: Essa rede pode aprender, armazenar e recuperar informações ao longo do tempo graças aos seus portões de entrada, esquecimento e saída, que controlam o fluxo de dados. Esses portões decidem quais informações manter ou descartar, tornando as LSTMs eficazes para capturar padrões de longo prazo em séries temporais. A arquitetura das LSTMs consiste em quatro camadas conectadas: uma principal que processa a entrada e o estado anterior, e três camadas de portões

controladas por funções logísticas. Os portões determinam quais partes do estado de longo prazo manter, apagar ou atualizar. Essa flexibilidade na manipulação do estado permite que as LSTMs tenham sucesso em diversas aplicações [8].

2) BLTSM: Variante que utiliza duas camadas LSTM, uma para processar na direção normal (*forward*) e outra na direção reversa (*backwar*d). Isso permite capturar informações contextuais tanto do passado quanto do futuro de cada ponto de dados na sequência, melhorando assim a capacidade de modelagem de dependências temporais em ambas as direções [8].

3) CRNN: Combina características das arquiteturas convolucionais e recorrentes, ao utilizar camadas convolucionais para extrair características espaciais da entrada e camadas recorrentes para modelar dependências temporais.

III. Configuração das Simulações

Esta seção descreve a configuração das simulações adotadas para avaliar o desempenho dos modelos propostos na atenuação de eco acústico em situações com conversação cruzada.

A. Caminho de Eco

O caminho de eco foi modelado por cinco respostas ao impulso de uma mesma sala, medidas com posições diferentes do alto-falante, disponíveis em [9]. As respostas ao impulso tiveram sua taxa de amostragem reduzida para 16kHz e em seguida foram truncadas em 1000 amostras.

Para não haver contaminação de dados, foi utilizado um $h(n)$ para o conjunto de teste, outro para o conjunto de validação e as três demais para o conjunto de treinamento. A Figura 2 exibe a resposta ao impulso utilizada no teste.

B. Sinais de Fala

A base TIMIT [10] foi utilizada para gerar os sinais $x(n)$, $s(n)$, $d(n)$ e $y(n)$ para os conjuntos de treinamento, validação e teste. Essa base contém 6.300 áudios, com uma sentença cada, gravados a uma taxa de amostragem de 16 kHz. Cada falante de 8 grandes regiões dialetais (RDs) dos Estados Unidos pronunciou 10 sentenças. Somente áudios com duração entre 2 e 5 s foram utilizados, totalizando 5.635 áudios.

Para gerar $x(n)$ foram utilizados 1.890 áudios. Para cada um dos 630 falantes, três áudios foram selecionados aleatoriamente e em seguida concatenados em todas as combinações de 3 em 3, gerando seis sinais de fala. Esse processo resultou no total de 3.780 sinais de fala, sendo 1.152 de falantes femininos e 2.628 de masculinos. Com o objetivo de formar pares de $x(n)$ e $s(n)$, e manter a mesma proporção de falantes de ambos os gêneros, foram descartados aleatoriamente alguns sinais gerados. Esse processo resultou no total de 2.304 sinais $x(n)$, metade de falantes femininos e metade de masculinos.

Para gerar $s(n)$ foram utilizados 2.304 dos áudios restantes, metade de falantes femininos e metade de masculinos. Essa quantidade foi escolhida para formar pares de $x(n)$ e $s(n)$, e manter a mesma proporção de falantes de ambos os gêneros. Nesses áudios foi realizado *zero-padding* aleatoriamente no início, no fim, ou igualmente no início e no fim, para não introduzir padrões previsíveis de conversação cruzada.

Os sinais $x(n)$ e $s(n)$ foram divididos em conjuntos de treinamento, validação e teste com base nas RDs: RDs 1, 3, 5, 6, 7 e 8 para treinamento; RD 2 para validação; RD 4 para teste. Dentro de cada conjunto, cada sinal $x(n)$ formou par com 5 sinais $s(n)$ escolhidos aleatoriamente. Os conjuntos de dados foram então reduzidos para ter 3.500 sinais no treinamento, 350 sinais na validação e 350 no teste, que são equivalentes a 92,7, 0,92 e 0,91 horas de áudio. A redução foi feita de forma que os pares tenham as seguintes proporções entre os gêneros dos locutores: 20% de $x(n)$ femininos e $s(n)$ masculinos; 20% de $x(n)$ masculinos e $s(n)$ femininos; 30% de $x(n)$ e $s(n)$ femininos; 30% de $x(n)$ e $s(n)$ masculinos. A distribuição dos dados é resumida na Tabela I.

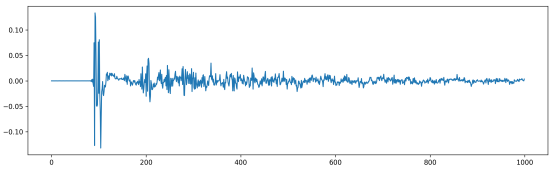

Fig. 2. Resposta ao impulso utilizada no conjunto de teste.

TABELA I
DISTRIBUIÇÃO DOS DADOS NO TREINAMENTO, VALIDAÇÃO E TESTE EM HORAS DE ÁUDIO.

$d(n) + s(n)$	Treinamento	Validação	Teste
30% M + 30% M	27,9	0,28	0,28
30% F + 30% F	27,8	0,28	0,27
20% M + 20% F	18,5	0,18	0,18
20% F + 20% M	18,5	0,18	0,18
Total	92,7	0,92	0,91

Nos conjuntos de treinamento e validação, os sinais $y(n)$ foram criados ao somar $d(n)$ e $s(n)$ com razões sinal-eco (SER, do inglês *signal-to-echo ratio*) escolhidas aleatoriamente entre $\{-6, -3, 0, 3, 6\}$ dB. Três conjuntos de teste de $y(n)$ foram criados: um com SER = 0 dB, outro com SER = 3,5 dB e outro com SER = 7 dB. Cada sinal $y(n)$ foi normalizado pelo seu maior valor absoluto, o qual também foi utilizado para normalizar os sinais $x(n)$, $d(n)$ e $s(n)$ que o geraram.

C. Modelos de Aprendizagem

Arquiteturas de quatro modelos foram empregadas. Elas foram treinadas por 20 épocas utilizando a GPU Tesla V100 do *Colaboratory*, o otimizador *Adamax*, o erro quadrático médio como função custo, taxa de aprendizado de 0,0003, e tamanho de lote de 256. Além disso, foram utilizados *callbacks* do *TensorFlow/Keras* para monitorar e controlar o treinamento dos modelos, como *EarlyStopping* e *ModelCheckpoint*.

1) LSTM: Modelo composto por quatro camadas LSTM com 300 unidades cada, configuradas para retornar sequências em todas as camadas. A camada de saída foi do tipo *TimeDistributed*, que aplica uma camada densa com uma unidade de saída e ativação sigmoide a cada etapa temporal da sequência, visto que a IRM possui valores no intervalo de 0 a 1.

2) BLSTM: Semelhante ao proposto em [3], esse modelo é composto por quatro camadas BLSTM com 300 unidades cada, configuradas para retornar sequências em todas as camadas. A camada de saída é uma camada *TimeDistributed*, que aplica uma camada densa com uma unidade de saída e ativação sigmoide a cada etapa temporal da sequência.

3) CRNN-1: Semelhante ao proposto em [11], esse modelo é composto pela combinação de camadas convolucionais, camadas LSTM, camadas convolucionais transpostas, e camadas

densas. No início, há quatro camadas convolucionais 1D com 64 filtros de tamanho 3 e função de ativação ReLu. Em seguida, duas camadas LSTM com 450 neurônios, configuradas para retornar sequências em todas as camadas. Posteriormente, quatro camadas convolucionais 1D transpostas com 64 filtros de tamanho 3 e função de ativação ReLu. Após isso, duas camadas densas, sendo a última com 161 unidades. Por fim, na saída, uma camada convolucional 1D com 1 filtro de tamanho 1 e função de ativação sigmoide.

4) CRNN-2: Composto pela combinação de camadas *Convolutional Long Short-Term Memory* de uma dimensão (ConvLSTM1D), camadas *Gated Recurrent Unit* (GRU), camadas convolucionais transpostas, e camadas densas. No início, há quatro camadas ConvLSTM1D, cada uma com 64 filtros de tamanho 3 e função de ativação ReLu. Em seguida, três camadas GRU com 450 neurônios, configuradas para retornar sequências em todas as camadas. Posteriormente, quatro camadas convolucionais transpostas 1D com 64 filtros de tamanho 3 e função de ativação ReLu. Após isso, uma camada densa para processar as informações das camadas convolucionais transpostas e outra camada densa com 161 unidades. Por fim, na saída, uma camada *TimeDistributed* com ativação sigmoide.

D. Aplicação da IRM

As STFTs $Y(k,\lambda)$, $X(k,\lambda)$, $S(k,\lambda)$ e $D(k,\lambda)$ foram calculadas utilizando uma janela de Hamming com duração de 20 ms, sobreposição de 50%, e uma transformada rápida de Fourier de 320 pontos. Para cada janela λ, $\hat{\xi}(k,\lambda)$ foi obtida conforme (6) e utilizada para computar $M(k,\lambda)$ conforme (5).

A máscara $M(k,\lambda)$ foi então aplicada a $Y(k,\lambda)$ conforme (2), obtendo $\hat{S}(k,\lambda)$. Por fim, a estimativa $\hat{s}(n)$ foi construída utilizando as STFTs inversas de $\hat{S}(k,\lambda)$ e *overlap-and-add*.

E. Métricas de Avaliação

1) W-PESQ: O W-PESQ (*Wideband Perceptual Evaluation of Speech Quality*) é um algoritmo para avaliação objetiva da qualidade de sinais de fala amostrados a 16 kHz [12], [13]. Ele compara representações psicoacústicas de um sinal de fala possivelmente degradado e sua referência não corrompida [14].

A pontuação bruta do W-PESQ pode ser mapeada para a escala 1-5 da opinião média (MOS, do inglês *Mean Opinion Score*), resultando na pontuação MOS-LQO (*MOS-Listening Quality Objective*) [15]. A correspondência entre essa escala e a classificação da categoria de degradação (DCR, do inglês *Degradation Category Rating*) é mostrada na Tabela II. No entanto, o máximo MOS-LQO fornecido pelo W-PESQ é 4,644, quando os sinais de referência e degradados são idênticos.

Neste trabalho, o algoritmo W-PESQ foi utilizado para avaliar o desempenho dos modelos quanto à qualidade sonora de $\hat{s}(n)$. Para isso, os trechos de $s(n)$ e $\hat{s}(n)$ nos períodos de conversação cruzada foram utilizados como os sinais de referência e degradado, respectivamente.

2) STOI: O STOI (*Short-Time Objective Intelligibility*) é uma métrica para avaliação objetiva da inteligibilidade de sinais de fala [16]. Ela é baseada em um coeficiente de correlação entre os envelopes temporais da fala limpa e degradada, em segmentos de curta duração (384 ms) e sobrepostos, fornecendo uma pontuação que varia de 0 a 1, onde pontuações mais elevadas indicam uma melhor inteligibilidade.

TABELA II
CORRESPONDÊNCIA ENTRE MOS-LQO E CATEGORIA DCR.

Pontuação	Categoria de Degradação
5	Inaudível
4	Audível, mas não incômoda
3	Pouco incômoda
2	Incômoda
1	Muito incômoda

Neste trabalho, o algoritmo STOI foi utilizado para avaliar o desempenho dos modelos quanto à inteligibilidade de $\hat{s}(n)$. Para isso, os trechos de $s(n)$ e $\hat{s}(n)$ nos períodos de conversação cruzada foram utilizados como os sinais de referência e degradado, respectivamente.

3) ERLE: O ERLE (*Echo Return Loss Enhancement*) é uma métrica para mensurar a atenuação do eco [17]. Ele é definido como a diferença instantânea, na escala decibel, entre as potências do eco acústico, $d(n)$, e do eco residual, $\hat{s}(n) - s(n)$. O ERLE será calculado apenas para períodos de fala única, i.e., quando $s(n) = 0$, por isso é definido como

$$\text{ERLE} = 10\log_{10}\left\{\frac{\text{E}[y^2(n)]}{\text{E}[\hat{s}^2(n)]}\right\} \tag{7}$$

Seus valores foram limitados ao intervalo de $[-100, 100]$ dB para evitar que ERLE $\to \infty$ quando não houver eco residual ou ERLE $\to -\infty$ quando não houver fala contaminada.

IV. RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados dos quatro modelos avaliados. Para comparação, são apresentados os resultados obtidos com a IRM ideal e sem máscara, estabelecendo assim os desempenhos máximo e mínimo para o problema.

Os valores médios de MOS-LQO, STOI e ERLE são apresentados na Tabela III. Observa-se que, em comparação ao caso sem processamento, os quatro modelos propostos proporcionam atenuação do eco acústico em todos os níveis de SER, melhorando assim a qualidade e inteligibilidade do sinal de eco compensado. Porém, exceto na inteligibilidade, o desempenho dos modelos propostos ainda está distante do caso ideal, principalmente em SER = 0 dB. Essa exceção se deve a alta inteligibilidade do sinal de fala contaminado.

Entre os modelos propostos, o BLSTM apresentou o melhor desempenho em todos os níveis de SER. A atenuação do eco gerada pelo BLSTM foi muito superior às dos outros modelos, superando em 166% a obtida com o LSTM para SER = 0 dB. No entanto, essa maior eficácia na atenuação de eco não se traduziu em grandes diferenças na qualidade e inteligibilidade do sinal de eco compensado. O modelo CRNN-2 também se destacou na atenuação do eco ao superar em 24% e 66% os modelos CRNN-1 e LSTM, respectivamente, para SER = 0 dB. A Figura 3 ilustra sinais processados pelo BLSTM e CRNN-2.

Em comparação ao estado-da-arte, os resultados alcançados pelo modelo proposto de BLSTM foram similares, embora ligeiramente inferiores, aos obtidos em [3]. Deve-se ressaltar que uma comparação direta não é adequada porque, além dos modelos BLSTM não serem idênticos, a manipulação da base TIMIT para gerar os sinais de fala e as respostas ao impulso utilizadas para o caminho acústico foram diferentes.

Na Tabela IV são apresentados os tempos necessários para treinamento e para inferência da IRM, assim como o atraso

TABELA III
RESULTADOS MÉDIOS DE MOS-LQO, STOI E ERLE.

Métrica	Modelo	SER (dB)		
		0	3,5	7
MOS-LQO	IRM Ideal	3,38	3,57	3,75
	BLSTM	2,06	2,46	2,81
	LSTM	1,86	2,21	2,51
	CRNN-1	1,92	2,29	2,58
	CRNN-2	1,90	2,24	2,51
	———	1,59	1,78	2,00
STOI	IRM Ideal	0,95	0,96	0,97
	BLSTM	0,85	0,90	0,93
	LSTM	0,80	0,86	0,89
	CRNN-1	0,82	0,88	0,91
	CRNN-2	0,82	0,87	0,89
	———	0,71	0,79	0,85
ERLE	IRM Ideal	99,50	99,50	99,50
	BLSTM	34,34	59,08	69,95
	LSTM	12,88	22,64	33,73
	CRNN-1	16,63	27,26	41,18
	CRNN-2	20,76	34,41	49,70
	———	0,00	0,00	0,00

TABELA IV
CUSTO COMPUTACIONAL DOS MODELOS.

Modelo	Tempo		
	Treinamento	Inferência	Atraso
BLSTM	19h31m38s	04m44s	63,7ms/s
LSTM	06h55m09s	01m48s	16,5ms/s
CRNN-1	06h43m17s	01m25s	15,8ms/s
CRNN-2	07h03m45s	02m05s	16,7ms/s

gerado no processamento. Esses tempos foram registrados utilizando a GPU T4 disponível no *Colaboratory*. Nota-se que o BLSTM possui um custo computacional muito mais elevado que os outros modelos, requerendo um tempo para o treinamento aproximadamente três vezes maior e gerando um atraso de processamento aproximadamente quatro vezes maior. Além disso, é fundamental destacar que o modelo BLSTM é não-causal, o que limita sua adequação para aplicações em tempo real [11]. Os outros modelos apresentaram custos computacionais semelhantes.

V. CONCLUSÕES

Este artigo avaliou a aplicação de redes neurais recorrentes e máscaras tempo-frequência na supressão de eco acústico. Quatro arquiteturas de redes neurais foram propostas para estimar a máscara de razão ideal, a qual foi utilizada para suprimir o eco em situações de conversação cruzada.

Os resultados indicaram que, nos momentos de fala única, modelos causais e de baixo custo computacional conseguem atenuar o eco em 20, 34 e 49 dB para razões sinal-eco de 0, 3,5 e 5 dB, respectivamente, enquanto modelos não-causais e de alto custo computacional alcançam resultados pelo menos 15 dB superiores. Porém, o melhor desempenho dos modelos de alto custo não se reflete significativamente na qualidade e na inteligibilidade dos trechos de conversação cruzada.

Portanto, a pesquisa traz luz à importância de, ao avaliar modelos de supressão de eco acústico, considerar a qualidade e inteligibilidade nos momentos de conversação cruzada. Essa

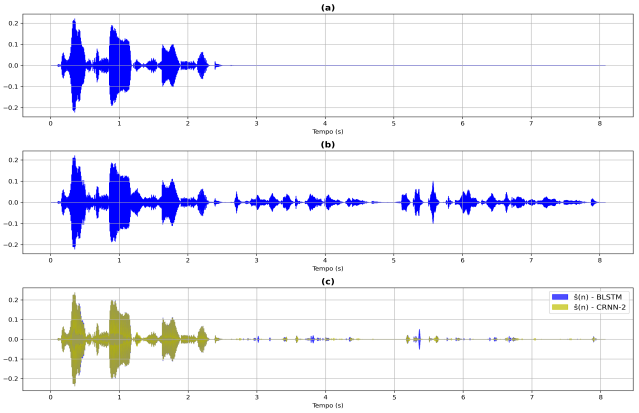

Fig. 3. Exemplo de sinais: (a) $s(n)$; (b) $y(n)$; (c) $\hat{s}(n)$.

análise pode auxiliar na escolha do modelo mais adequado, levando em consideração não apenas o desempenho técnico, mas também as necessidades práticas e operacionais.

REFERÊNCIAS

[1] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*. John Wiley & Sons, 2008.
[2] Z. Wang, Y. Na, B. Tian, and Q. Fu, "NN3A: neural network supported acoustic echo cancellation, noise suppression and automatic gain control for real-time communications," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 661–665.
[3] H. Zhang and D. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Interspeech*, 2018, pp. 3239–3243.
[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.
[5] R. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.
[6] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.
[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] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc, 2022.
[9] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *International Conference on Digital Signal Processing*, 2009, pp. 1–5.
[10] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," 1989.
[11] H. Zhang, "Deep learning for acoustic echo cancellation and active noise control," Ph.D. dissertation, The Ohio State University, 2022.
[12] ITU-T, "Perceptual evaluation of speech quality (PESQ): objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs," International Telecommunications Union, Geneva, Switzerland, 2001.
[13] ——, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunications Union, Geneva, Switzerland, 2005.
[14] B. C. Bispo, P. A. A. Esquef, L. W. P. Biscainho, A. A. de Lima, F. P. Freeland, R. A. de Jesus, A. Said, B. Lee, R. W. Schafer, and T. Kaller, "EW-PESQ: a quality assessment method for speech signals sampled at 48 kHz," *Journal of the Audio Engineering Society*, vol. 58, no. 4, pp. 251–268, April 2010.
[15] ITU-T, "Mean opinion score (mos) terminology," International Telecommunications Union, Geneva, Switzerland, 2006.
[16] 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.
[17] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic press library in signal processing*. Elsevier, 2014, vol. 4, pp. 807–877.