Impacto da Acurácia de Classificação de Métodos de Decodificação da Atenção Auditiva no Desempenho do Filtro de Wiener Multicanal

Lucian S. Schiavon, Márcio H. Costa e José C. M. Bermudez

***Resumo*— Este trabalho apresenta uma investigação sobre o impacto da acurácia de classificação e da duração da janela de decisão de métodos de decodificação de atenção auditiva (AAD) sobre o desempenho do filtro de Wiener multicanal. Simulações computacionais em cenários acústicos compostos por dois locutores simultâneos resultaram em evidências de que acurácias de classificação maiores que 85% são suficientes para propiciar inteligibilidade adequada para conversação. Foram propostas duas formas de simulação de erros de classificação e utilizadas métricas objetivas para estimação de inteligibilidade e qualidade. Os resultados obtidos estabelecem um ponto de partida para auxiliar a escolha do método AAD em aparelhos auditivos.**

Palavras-Chave*— *Decodificação de atenção auditiva, Aparelho auditivo, Inteligibilidade da fala.

***Abstract*— This paper investigates the impact of classification accuracy and length of the decision window of auditory attention decoding (AAD) methods on the performance of the multichannel Wiener filter. Computational simulations in acoustic scenarios composed of two simultaneous speakers resulted in evidence that classification accuracies greater than 85% are sufficient to provide adequate intelligibility for conversation. Two methods for simulating classification errors were proposed, and objective metrics were used to estimate intelligibility and quality. The results establish a starting point to assist in choosing the AAD method in hearing aids.**

Keywords*—*Auditory attention decoding, Hearing aids, Speech intelligibility.

I. INTRODUÇÃO

Segundo a Organização Mundial de Saúde, cerca de 430 milhões de pessoas no mundo sofrem algum tipo de perda auditiva. Estima-se que, em 2050, aproximadamente 2,5 bilhões de pessoas terão algum grau de perda auditiva e que 700 milhões de pessoas necessitarão de reabilitação [1].

Pessoas com perda auditiva grave estão sujeitas a sérios problemas sociais, que podem resultar em limitações profissionais e isolamento social [2]. Aparelhos auditivos e implantes cocleares são os dispositivos mais comuns na compensação de perdas auditivas. No entanto, ainda existem diversos desafios na utilização dos mesmos, principalmente em situações em que várias pessoas falam simultaneamente ou em ambientes com ruído.

No contexto da aplicação em aparelhos auditivos e implantes cocleares, existem diversos métodos para realizar a separação de uma mistura de sinais de áudio provenientes de diferentes fontes acústicas [3]. Em geral, o objetivo dessa separação é permitir a enfatização de um sinal específico, sendo os demais considerados como ruído. No entanto, um problema recorrente nesse tipo de problema é a identificação do sinal de interesse do ouvinte.

Uma recente abordagem para essa questão é a extração da informação de atenção auditiva do ouvinte a partir do eletroencefalograma (EEG). Os métodos utilizados para essa tarefa são chamados na literatura de decodificação da atenção auditiva (AAD – *Auditory Attention Decoding*) [3].

Métodos AAD utilizam a informação de correlação entre a envoltória do EEG e dos sinais acústicos captados pelos microfones do aparelho auditivo para identificar a fonte acústica ou sinal de interesse do usuário em diferentes instantes de tempo, cuja duração é chamada de janela de decisão.

O método de redução de ruído mais estudado para aplicações em aparelhos auditivos é o filtro de Wiener multicanal (MWF – *Multichannel Wiener Filter*). Em geral, o principal fator limitante de seu desempenho é a estimação das estatísticas de segunda ordem do sinal desejado [4]. A forma convencional considera que fala e ruído são não correlacionados e utiliza um detector de fala (VAD – *Voice Activity Detector*). Entretanto, os resultados do VAD não são adequados quando a fonte interferente é fala humana.

A maioria dos trabalhos na área de AAD se concentra no desempenho do processo de classificação em cenários com dois locutores [3]. No entanto, não se encontram na literatura estudos sobre o impacto da acurácia de classificação sobre o desempenho de métodos de redução de ruído (como o MWF), em especial, em termos da qualidade e inteligibilidade dos sinais processados.

Este trabalho apresenta um estudo sobre o impacto da acurácia de classificação e da janela de decisão dos algoritmos AAD no desempenho do MWF em termos de critérios objetivos de qualidade e inteligibilidade da fala. O objetivo é a determinação de requisitos mínimos de acurácia de classificação e de duração da janela de decisão para obter níveis mínimos de inteligibilidade e qualidade para a compreensão da fala. Para tanto, um sistema de redução de ruído baseado no MWF e AAD [5] foi implementado em um cenário acústico composto por dois locutores simultâneos, com razão sinal-interferência (SIR – *Signal to Interference Ratio*) de 0 dB e ruído do ambiente com razão sinal ruído (SNR – *Signal to Noise Ratio*) de 20 dB.

O trabalho foi organizado da seguinte forma: na Seção II são apresentados a discussão do problema e o sistema de redução de ruído. Na Seção III é detalhada a metodologia utilizada, além

Lucian S. Schiavon, Márcio H. Costa e José C. M. Bermudez, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis-SC, E-mails: lucianschiavon@gmail.com; costa@eel.ufsc.br; j.bermudez@ieee.org.

dos parâmetros da implementação e os critérios objetivos empregados na avaliação. A Seção IV apresenta os resultados obtidos, os quais são discutidos na Seção V. As conclusões sobre o trabalho estão na Seção VI.

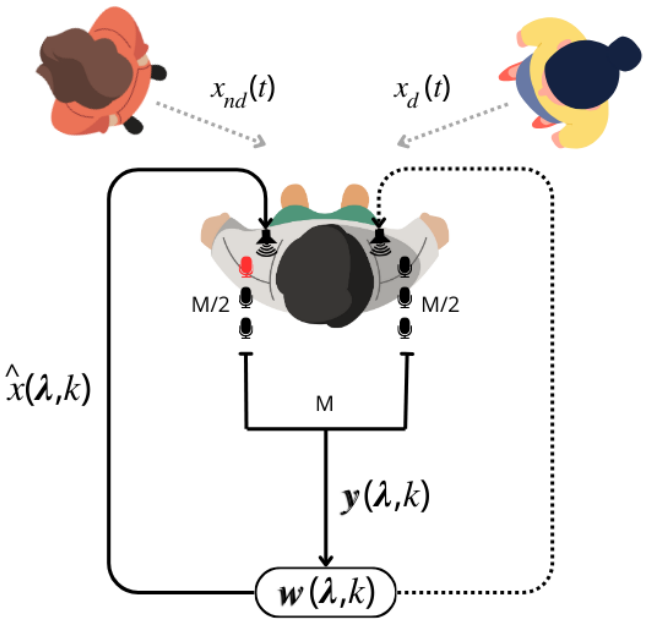

Fig. 1. Diagrama do cenário acústico analisado.

II. Descrição do Problema

Considere o cenário acústico mostrado na Fig. 1. Os microfones do aparelho auditivo captam uma mistura dos sinais de fala de N locutores, além do ruído presente no ambiente (campo acústico difuso e/ou ruído elétrico). O sinal $x_d(t)$ é o sinal desejado (sinal de fala de interesse do ouvinte) e o sinal $x_{nd}(t)$ é o sinal indesejado. Os sinais captados pelos microfones do aparelho auditivo são transformados para o domínio tempo-frequência, sendo representados por

$$y_m(\lambda,k) = \sum_{i=1}^{N} x_{m,i}(\lambda,k) + n_m(\lambda,k), \quad (1)$$

em que λ indica o *frame* de tempo e k representa o *bin* de frequência. O sinal de fala do i-ésimo locutor captado pelo m-ésimo microfone é representado por $x_{m,i}(\lambda,k)$ e o ruído ambiente por $n_m(\lambda,k)$. Os sinais de fala podem ser modelados por

$$x_{m,i}(\lambda,k) = a_{m,i}(\lambda,k)s_i(\lambda,k), \quad (2)$$

em que $s_i(\lambda,k)$ é o sinal de fala gerado pelo i-ésimo locutor e $a_{m,i}(\lambda,k)$ é a função de transferência (ATF – *Acoustic Transfer Function*) que modela o caminho de propagação acústica entre a fonte e o microfone que capta o som. Assume-se que a fala de interesse é definida por $s_1(\lambda,k)$.

A equação (1) pode ser representada no formato vetorial, de forma a conter todos os sinais captados pelos M microfones, de forma que

$$\mathbf{y}(\lambda,k) = \mathbf{a}_1(\lambda,k)s_1(\lambda,k) + \sum_{i=2}^{N} \mathbf{a}_i(\lambda,k)s_i(\lambda,k) + \mathbf{n}(\lambda,k) \quad (3)$$

em que $\mathbf{a}_i(\lambda,k) = [\, a_{1,i}(\lambda,k)\ a_{2,i}(\lambda,k)\ \ldots\ a_{M,i}(\lambda,k)\,]^T$ é o vetor de ATFs; $\mathbf{y}(\lambda,k) = [\, y_1(\lambda,k)\ y_2(\lambda,k)\ \ldots\ y_M(\lambda,k)\,]^T$ e $\mathbf{n}(\lambda,k) = [\, n_1(\lambda,k)\ n_2(\lambda,k)\ \ldots\ n_M(\lambda,k)\,]^T$ são, respectivamente, os vetores de fala contaminada e de ruído difuso. E, portanto,

$$\mathbf{y}(\lambda,k) = \mathbf{x}(\lambda,k) + \mathbf{r}(\lambda,k) \quad (4)$$

em que $\mathbf{x}(\lambda,k) = s_1(\lambda,k)[\, a_{1,1}(\lambda,k)\ \ a_{2,1}(\lambda,k)\ \ldots\ a_{M,1}(\lambda,k)\,]^T$ e $\mathbf{r}(\lambda,k) = [\, r_1(\lambda,k)\ \ r_2(\lambda,k)\ \ \ldots\ \ r_M(\lambda,k)\,]^T$ representam, respectivamente, os vetores de fala de interesse e de ruído aditivo.

O problema de redução de ruído consiste em estimar o sinal de fala desejado a partir dos sinais captados pelos M microfones do aparelho auditivo. A estimativa do sinal de fala desejado no microfone de referência é representada por

$$\hat{x}(\lambda,k) = \mathbf{w}^H(\lambda,k)\mathbf{y}(\lambda,k) \quad (5)$$

em que $(\cdot)^H$ representa a operação de transposição conjugada e $\mathbf{w}(\lambda,k) = [\, w_1(\lambda,k)\ w_2(\lambda,k)\ \ldots\ w_M(\lambda,k)\,]^T$ é o filtro de redução de ruído. Neste trabalho, o microfone frontal esquerdo, $m = 1$, foi escolhido como o microfone de referência (em vermelho na Fig. 1).

A. Filtro de Wiener multicanal

O MWF é uma técnica amplamente utilizada para a redução de ruído em aparelhos auditivos, resultando em uma estimação linear da fala desejada com o menor erro quadrático médio (MSE – *Mean Squared Error*) [4].

O vetor de coeficientes do MWF, $\mathbf{w}(\lambda,k)$, é determinado a partir do seguinte problema de minimização [4]

$$\mathbf{w}(\lambda,k) = \min_{\mathbf{w}} \mathrm{E}\{|x_1(\lambda,k) - \mathbf{w}^H\mathbf{y}(\lambda,k)|^2\} \quad (6)$$

que resulta em

$$\mathbf{w}^{\text{otimo}}(\lambda,k) = \boldsymbol{\Phi}_{\mathbf{yy}}^{-1}(\lambda,k)\boldsymbol{\Phi}_{\mathbf{xx}}(\lambda,k)\mathbf{q}, \quad (7)$$

em que $\boldsymbol{\Phi}_{\mathbf{yy}}(\lambda,k) = \mathrm{E}\{\mathbf{y}(\lambda,k)\mathbf{y}^H(\lambda,k)\}$, $\boldsymbol{\Phi}_{\mathbf{xx}}(\lambda,k) = \mathrm{E}\{\mathbf{x}(\lambda,k)\mathbf{x}^H(\lambda,k)\}$, $\mathbf{x}(\lambda,k) = \mathbf{a}_1(\lambda,k)s_1(\lambda,k)$ e $\mathbf{q} = [\, 1\ 0\ 0\ \ldots\ 0\,]^T$ é o vetor de seleção. A matriz $\boldsymbol{\Phi}_{\mathbf{yy}}(\lambda,k)$ é estimada diretamente do sinal captado, enquanto que $\boldsymbol{\Phi}_{\mathbf{xx}}(\lambda,k) = \boldsymbol{\Phi}_{\mathbf{yy}}(\lambda,k) - \boldsymbol{\Phi}_{\mathbf{rr}}(\lambda,k)$ em decorrência da consideração de não correlação entre fala e ruído.

O fator prático limitante de desempenho do MWF na aplicação em questão é a acurácia de estimação da matriz de coerência do sinal desejado, visto que este sinal se encontra misturado ao ruído aditivo que pode conter outras fontes de interferência.

B. Sistema de redução de ruído baseado em AAD

Em [5] foi proposto um sistema de redução de ruído baseado em AAD, cujo diagrama em blocos se encontra representado na Fig. 2. Esse sistema assume a existência de apenas duas fontes acústicas, sendo uma delas a fala de interesse e outra a fala interferente. Nesse sistema, os sinais captados pelos M microfones dos aparelhos auditivos são processados por um sistema de separação cega de fontes, gerando aproximações com detalhamento limitado para os sinais de fala de interesse (locutor L_1) e interferente (locutor L_2). Na saída do sistema de separação cega são utilizados VADs que identificam os trechos de fala e ruído. Dessa forma, é possível estimar a matriz de coerência $\boldsymbol{\Phi}_{\mathbf{rr}}(\lambda,k)$ através do método de *covariance whitening* [6]. Os filtros ótimos do MWF são calculados a partir dessas informações e então utilizados para filtrar os sinais captados pelos microfones para obter estimativas dos sinais de fala de cada um dos locutores. Finalmente, o algoritmo de AAD faz a seleção do sinal de fala que será reproduzido para o usuário do aparelho auditivo.

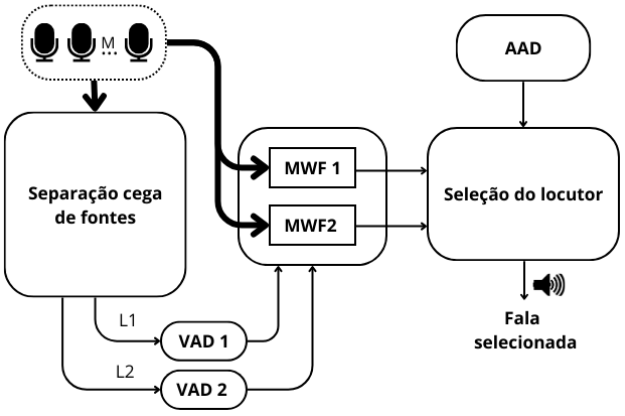

Fig. 2. Diagrama de blocos do sistema de redução de ruído baseado em AAD [5].

O sistema de separação cega de fontes pode ser implementado de várias formas, como, por exemplo, utilizando redes neurais profundas ou análise de componentes independentes (ICA – *Independent Component Analysis*) [5].

Uma abordagem comum para AAD utiliza um procedimento de decodificação, aplicado em todos os canais de EEG, para reconstruir a envoltória da fala do locutor-alvo enquanto bloqueia outras atividades neurais que não estão relacionadas ao problema [3]. Posteriormente, essa envoltória decodificada é correlacionada com a envoltória da fala de cada um dos locutores. Como resultado, o locutor alvo é escolhido como aquele que possui maior coeficiente de correlação de Pearson. Em geral, o resultado dos métodos AAD depende fortemente da duração da janela de decisão, uma vez que a estimativa do coeficiente de correlação de Pearson é ruidosa devido à baixa SNR dos sinais envolvidos [3].

Considerando a aplicação em aparelhos auditivos, é necessário que a duração dessa janela de decisão seja pequena o suficiente para não prejudicar a experiência do usuário. Em alguns trabalhos, foram obtidas acurácias em torno de 80 a 90% em um cenário com dois locutores, porém com janelas de decisão de aproximadamente 30s, as quais são consideradas longas em um contexto prático [7]. No entanto, em trabalhos utilizando redes neurais convolucionais e redes neurais profundas, foi alcançada a mesma faixa de acurácia com janelas de decisão de aproximadamente 1s, ou até menores [8] [9].

III. METODOLOGIA

Neste trabalho foi realizada uma avaliação do desempenho potencial da arquitetura de redução de ruído baseada em MWF e AAD apresentada na Fig. 2 para $N = 2$ locutores [5]. Os parâmetros utilizados para a avaliação foram a duração da janela de decisão e a acurácia da decisão sobre o locutor de interesse. Para cada combinação de acurácia e duração da janela de decisão foram realizadas 150 simulações computacionais realizadas em *MATLAB*^{TM}.

A. Cenário acústico

As simulações foram feitas a partir da emulação de um ambiente anecoico. Um áudio de conversação de múltiplos locutores (*babble*) em uma cafeteria [10] foi utilizado como ruído difuso de fundo com SNR de 20 dB em relação ao sinal de fala desejado. Segundo [11], uma SNR de 20 dB qualifica um ambiente como apropriado para conversação.

Foram utilizadas respostas ao impulso relacionadas com a cabeça (HRIR – *Head-related impulse response*) para seis microfones obtidas a partir de um aparelho auditivo biauricular, sendo três microfones no lado direito e três microfones no lado esquerdo ($M = 6$) [12]. O microfone frontal do lado esquerdo foi considerado como microfone de referência. O ângulo de 0º corresponde à posição frontal ao ouvinte. O locutor desejado está localizado 90º à direita do ouvinte, enquanto o locutor não-desejado está 90º à esquerda do ouvinte. A SIR entre os locutores foi definida como 0 dB.

B. Sinais de fala

Os sinais de fala dos dois locutores correspondem a duas pessoas do sexo feminino, oriundos da base de dados desenvolvida em [13]. Foram usados 20 áudios com trechos de aproximadamente 4s. A frequência de amostragem original de 11.025 Hz foi interpolada para 16 kHz.

C. Domínio tempo-frequência

A transformação do domínio de tempo discreto para o domínio tempo-frequência foi realizada a partir da transformada de Fourier de tempo curto (STFT – *Short Time Fourier Transform*) com uma janela de 512 amostras e sobreposição de 50%.

D. Separação cega de fontes

De forma a evitar a influência do sistema de separação cega sobre o desempenho do sistema de redução de ruído, assumiu-se que as falas dos dois locutores podem ser perfeitamente separadas, configurando um separador de fontes ideal. Foi também utilizado um VAD ideal, obtido através da observação dos sinais. Essas considerações permitem avaliar o impacto do sistema AAD sobre o máximo desempenho possível do MWF.

E. Decodificação de atenção auditiva

O impacto da acurácia de classificação do sistema AAD sobre o desempenho do MWF foi avaliado para valores de acurácia entre 70 e 90% com intervalos de 5%. A duração da janela de decisão foi analisada para valores de 1s, 0,5s e 0,25s [8] [9].

Duas formas foram utilizadas para simular a seleção do locutor de interesse. Ambas foram definidas como sorteios aleatórios realizados a partir de uma função de probabilidade binomial (locutor L_1 e L_2). No primeiro método, a probabilidade de seleção é fixa, definindo diretamente a acurácia desejada. No segundo, a probabilidade de acerto do locutor de interesse foi determinada pelo coeficiente de correlação entre as envoltórias dos sinais de fala desejada e interferente. Quanto maior for a correlação entre os sinais maior é a probabilidade de erro de seleção do locutor desejado.

F. Critérios objetivos de qualidade e inteligibilidade da fala

Diferentemente dos trabalhos apresentados na literatura, que se concentram na acurácia de seleção do AAD, neste trabalho foram utilizados o PESQ (*Perceptual Evaluation of Speech Quality*) de banda larga [14] e o STOI (*Short-time Objective Intelligibility measure*) [15] para avaliação da qualidade e inteligibilidade da fala, respectivamente.

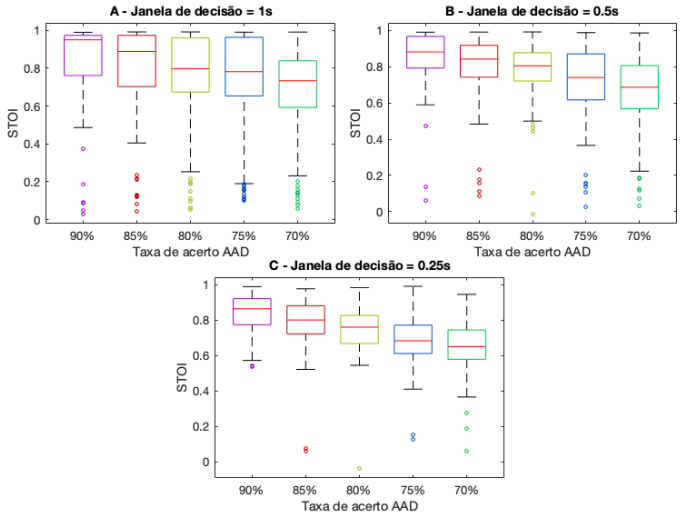

Fig. 3. STOI para diferentes janelas de decisão e acurácia de seleção do AAD. Método de seleção aleatória com probabilidade fixa.

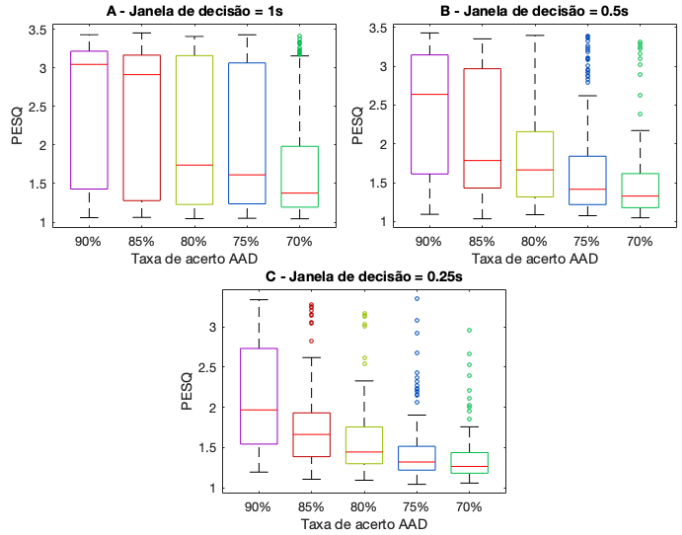

Fig. 4. PESQ para diferentes janelas de decisão e acurácia de seleção do AAD. Método de seleção aleatória.

IV. Resultados

A Fig. 3 e a Fig. 4 mostram diagramas de caixa dos valores de STOI e PESQ, respectivamente, para diferentes janelas de decisão e acurácia de seleção, considerando-se que a seleção do locutor foi realizada de forma aleatória com probabilidade fixa.

A Fig. 5 e a Fig. 6 também apresentam diagramas de caixa dos valores de STOI e PESQ. No entanto, nessas situações, a seleção do locutor foi realizada a partir do coeficiente de correlação entre as envoltórias dos sinais de fala desejada e interferente.

Ao observar os valores das medianas na Fig. 3 e na Fig. 5, pode-se notar um comportamento aproximadamente linear do STOI em função da taxa de acerto, independentemente da duração da janela de decisão. Comportamento semelhante também é verificado para o PESQ através da Fig. 4 e da Fig. 6. No entanto, para ambos os métodos, é possível verificar que, no caso específico da janela de decisão de 1s houve um decaimento significativo nos valores medianos do PESQ para acurácias menores que 85%.

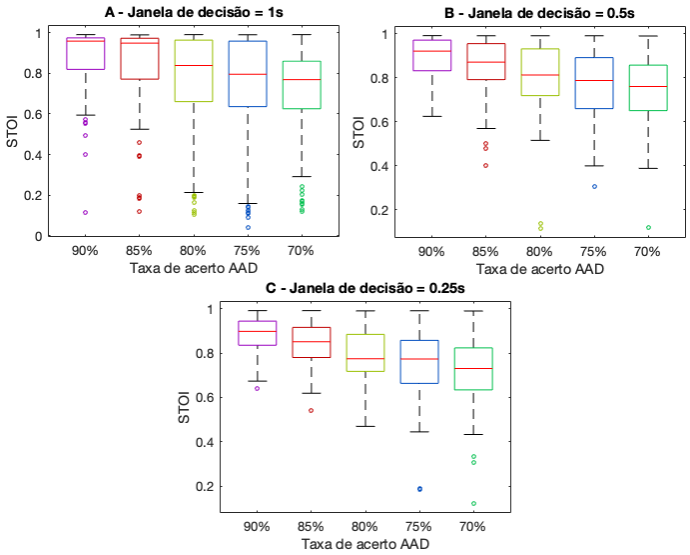

Fig. 5. STOI para diferentes janelas de decisão e acurácia de seleção do AAD. Método de seleção aleatória com probabilidade baseada no coeficiente de correlação entre as envoltórias do sinal desejado e interferente.

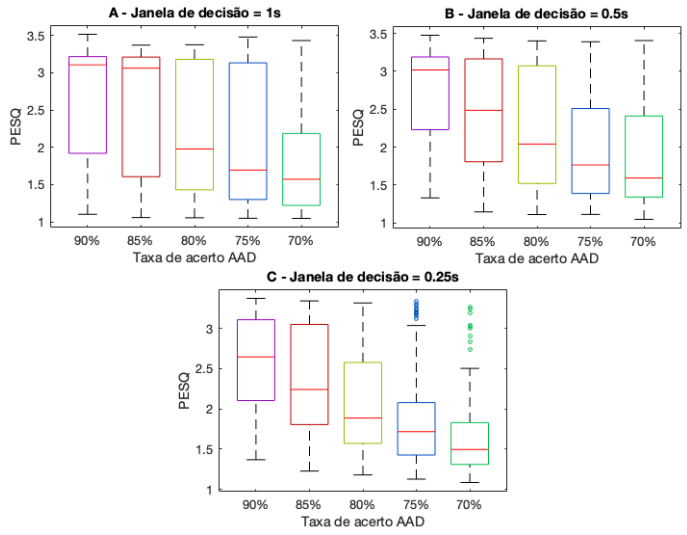

Fig. 6. PESQ para diferentes janelas de decisão e acurácia de seleção do AAD. Método de seleção aleatória com probabilidade baseada no coeficiente de correlação entre as envoltórias do sinal desejado e interferente.

V. Discussão

Apesar dos métodos utilizados para a simulação do processo de seleção de atenção serem artificiais, os resultados obtidos apresentam coerência com o desempenho esperado. Embora distintos em suas concepções probabilísticas, apresentam resultados congruentes. O método de seleção baseado na correlação entre envoltórias apresenta um suporte teórico associado ao funcionamento de métodos AAD reais e, portanto, é assumido como o de melhor representação [3].

Segundo [16], valores de STOI a partir de 0,75 correspondem a uma boa inteligibilidade. Mesmo para outras

métricas de inteligibilidade, como os valores do índice de transmissão da fala (STI – *Speech Transmission Index*) ou os valores do índice de inteligibilidade da fala (SII – *Speech Intelligibility Index*), valores maiores que 0,75 caracterizam condições de inteligibilidade boas [17] [18], considerando que ambas as métricas também geram resultados em um intervalo entre 0 e 1. Portanto, considera-se neste trabalho que valores de STOI maiores que 0,75 representam situações de inteligibilidade adequadas para conversação. Analisando os valores de mediana nos diagramas de caixas da Fig. 5, observa-se que o único valor de STOI abaixo de 0,75 refere-se à acurácia de 70% para uma janela de decisão de 0,25s. No entanto, se utilizarmos como referência o primeiro quartil, apenas taxas de acerto acima de 85% apresentam valores de STOI acima de 0,75, considerando todos os tamanhos de janela.

De acordo com [19], o valor de PESQ igual a 1,5 é considerado o mínimo valor discriminável de qualidade, enquanto que 0,2 é uma variação perceptível para o ouvinte. Observando os valores de mediana nos diagramas da Fig. 6, verifica-se que o único valor abaixo de 1,5 é referente à acurácia de 70% para janela de decisão de 0,25s. Entretanto, ao empregar como referência o primeiro quartil, são encontrados valores de PESQ acima de 1,5 somente a partir de taxas de acerto de 85%, para todos os tamanhos de janela. É interessante notar também que há mudanças perceptíveis de PESQ (maiores que 0,2) nos valores do primeiro quartil entre as taxas de acerto de 90% e 85%, com exceção da janela de decisão de 1s. Ou seja, em termos de qualidade da fala, não parece haver diferenças para acurácias de 85% a 90% na etapa de AAD.

A partir dos resultados obtidos, pode-se inferir que existem evidências de que acurácias de classificação do AAD a partir de 85%, para os três tamanhos investigados de janelas de decisão, resultam em inteligibilidade adequada para conversação. Essa constatação é de grande interesse para o projeto de sistemas AAD visto que esses sistemas podem incorporar altos custos computacionais ao aparelho auditivo, como é o caso de métodos baseados em redes neurais convolucionais [8] [9] ou redes neurais totalmente conectadas [20].

Os resultados obtidos neste trabalho são um ponto de partida para o auxílio à tomada de decisão sobre o modelo de decodificação de atenção auditiva a ser utilizado, considerando o impacto sobre a inteligibilidade e o custo computacional.

VI. CONCLUSÕES

Neste trabalho foi investigado o impacto da acurácia de classificação e da janela de decisão de métodos de decodificação de atenção auditiva sobre o desempenho do filtro de Wiener multicanal. Simulações computacionais em cenários acústicos compostos por dois locutores simultâneos resultaram em evidências de que acurácias de classificação maiores que 85% são suficientes para suprir inteligibilidade adequada, independentemente da duração da janela de decisão.

REFERÊNCIAS

[1] Organização Mundial de Saúde (OMS). “Deafness and hearing loss”. Abril 2021. Disponível em: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] R. Einhorn. “Hearing aid technology for the 21st Century: A proposal for universal wireless connectivity and improved sound quality”, *IEEE Pulse*, v. 8, n. 2, p. 25-28, 2017.

[3] S. Geirnaert et al., “Electroencephalography-based auditory attention decoding toward neurosteered hearing devices”, *IEEE Signal Processing Magazine*, v. 38, n. 4, p. 89-102, 2021.

[4] B. Cornelis et al., “Theoretical analysis of binaural multimicrophone noise reduction techniques”, *IEEE Transactions on Audio, Speech, and Language Processing*, v. 18, n. 2, p. 342–355, 2009.

[5] N. Das et al., “Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding”, *Journal of Neural Engineering*, v. 17, n.4, p. 046039, 2020.

[6] R. Serizel et al., “Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 22, n. 4, p. 785-799, 2014.

[7] W. Biesmans, N. Das, T. Francart, e A. Bertrand, “Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, v. 25, n. 5, p. 402-412, 2017.

[8] S. Vandecappelle et al., “EEG-based detection of the locus of auditory attention with convolutional neural networks”, *Neuroscience*, p. 1-17, 2021.

[9] Z. Xu et al., “Decoding selective auditory attention with EEG using a transformer model”, *Methods*, v. 204, p. 410-417, 2022.

[10] P. C. Loizou, “Speech Enhancement: Theory and Practice”, 2ed., Taylor and Francis, London, 2013. DVD-ROM.

[11] K. Smeds, F. Wolters e M. Rung, “Estimation of signal-to-noise ratios in realistic sound scenarios”, *Journal of the American Academy of Audiology*, v. 26, p. 183-196, 2015.

[12] H. Kayser et al. “Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses”. In: *EURASIP Journal on Advances in Signal Processing*, 2009.

[13] C. A. Ynoguti, “Reconhecimento de fala contínua usando modelos ocultos de Markov”. Ph. D. Universidade Estadual de Campinas, 1999.

[14] ITU, Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation P. 862. 2000.

[15] C. H. Taal et al., “A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech”, *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Dallas, United States, pp. 4214-4217, 2010.

[16] S. Graetzer e C. Hopkins, “Intelligibility prediction for speech mixed with white Gaussian noise at low signal-to-noise ratios”, *J. Acoust. Soc. Am.* 1, 2021.

[17] ISO 9921: Ergonomics - Assessment of speech communication. 2003.

[18] ANSI S3.5-1997: American national standard - methods for calculation of the speech intelligibility index (SII). American National Standards Institute & Inc., New York, 1997.

[19] A. Servetti, e J. C. De Martin, “802.11 MAC protocol with selective error detection for speech transmission”, *Lecture Notes in Computer Science*, p. 509-519, 2005.

[20] T. de Taillez, B. Kollmeier, e B. T. Meyer, “Machine learning for decoding listeners’ attention from electroencephalography evoked by continuous speech”, *European Journal of Neuroscience*, v. 51, n. 5, p. 1234-1241, 2020.