# Sobre a influência de distorções acústicas na classificação de patologias laríngeas

Vinícius J. D. Vieira, Rafael R. Pertum e Renato Candido

***Resumo*— Neste trabalho, é realizada uma análise sobre a influência de variações acústicas na classificação de patologias laríngeas. Para tanto, é observado o comportamento das seguintes características extraídas dos sinais de voz: MFCC (*Mel-Frequency Cepstral Coefficients*), LPC (*Linear Predictive Coefficients*), $F0$ (frequência fundamental), *Jitter* e *Shimmer*. As variações acústicas consideradas são: diferentes taxas de amostragem, diferentes níveis de ruído, e diferentes níveis de reverberação. As patologias consideradas são: edema de Reinke, carcinoma, leucoplasia, pólipos e paralisia nas pregas vocais. A etapa de classificação é realizada com análise discriminante quadrática. Os resultados indicam que as medidas apresentam diferentes comportamentos de acurácia de acordo com a distorção acústica. A medida MFCC foi a que apresentou maior robustez em todos os cenários de variação, considerando um desempenho acima de 70% de acurácia, sensibilidade e especificidade.**

***Palavras-Chave*— Processamento de sinais de voz, classificação de patologias laríngeas, distorções acústicas, características acústicas**

***Abstract*— In this work, an analysis is carried out on the influence of acoustic variations in the classification of laryngeal pathologies. In this context, it is observed the behavior of the following characteristics extracted from voice signals: MFCC (Mel-Frequency Cepstral Coefficients), LPC (Linear Predictive Coefficients), $F0$ (fundamental frequency), Jitter and Shimmer. The acoustic variations considered are: different sampling rates, different noise levels, and different reverberation levels. The pathologies considered are: Reinke's edema, carcinoma, leukoplakia, polyps and vocal fold paralysis. The classification stage is performed with quadratic discriminant analysis. The results indicate that the measurements present different accuracy behaviors according to the acoustic distortion. The MFCC measure was the one that presented the greatest robustness in all variation scenarios, considering a performance above 70% in terms of accuracy, sensitivity and specificity.**

***Keywords*— Speech signal processing, laryngeal pathologies classification, acoustic distortions, acoustic characteristics**

## I. Introdução

A voz é o principal meio de comunicação do ser humano. Por meio dela, é possível identificar informações como identidade do locutor, conteúdo da fala, estado emocional e saúde vocal. Este último é tópico de diversas pesquisas nas últimas décadas [1–3]. A presença de distúrbios vocais pode impactar significativamente a qualidade de vida e a eficácia da comunicação [4]. Portanto, a triagem, o diagnóstico e o acompanhamento do tratamento vocal se torna uma prática importante para pessoas de todas as faixas etárias. Entre os impactos consequentes da pandemia da COVID-19, pode ser citada a dificuldade de realizar consultas presenciais por parte dos pacientes. Como consequência, surgiu uma oportunidade de ampliar o raio geográfico de atendimento por meio da telessaúde, principalmente para quem mora longe dos grandes centros médicos [5], [6].

Vinícius J. D. Vieira, Rafael R. Pertum e Renato Candido, Depto. de Processamento de Voz, SiDi, Alphaville, Campinas, SP - Brasil, e-mails: {vinicius.dv, r.pertum, r.candido}@sidi.org.br.

Por meio da telessaúde, o paciente pode ser atendido diretamente de casa, utilizando um *smartphone*, por exemplo. Em contrapartida, a análise vocal enfrenta desafios adicionais devido às variações acústicas ambientais e às características dos dispositivos de gravação. Essas variações podem incluir diferentes taxas de amostragem no processamento do dispositivo, níveis de ruído e reverberação, que afetam a precisão das medições acústicas e, consequentemente, a qualidade do diagnóstico. A implementação de telessaúde na avaliação vocal permite uma maior acessibilidade e conveniência para os pacientes, mas exige métodos robustos para lidar com essas variações acústicas.

Na literatura ainda há poucos trabalhos que relacionam variações acústicas ambientais em um contexto de telessaúde para avaliação de distúrbios vocais. Em [7], foi investigado o uso de *smartphones* na obtenção de sinais de voz em ambientes ruidosos, sendo constatado que as características utilizadas, *Jitter* e razão ruído-harmônicos, sofrem com os efeitos das distorções ruidosas. Em [8], é investigada a influência da variação de dispositivos de gravação na extração de medidas acústicas para avaliação da saúde vocal, sendo observado que a frequência fundamental ($F0$) proporciona menos erro de classificação que *Jitter* e *Shimmer*, por exemplo. Em [9], técnicas de aprendizado de máquina foram empregadas para classificar sinais de vozes saudáveis e patológicas captadas por *smartphone*, utilizando como característica a medida MFCC, chegando a uma acurácia de 98,3% com classificador baseado em k-*nearest neighbors*.

Neste trabalho, é investigado o impacto de distorções acústicas em características extraídas dos sinais de voz para classificação de patologias laríngeas. A principal contribuição deste estudo é apresentar experimentos realizados em três diferentes cenários possíveis com o advento do uso da telessaúde na prática clínica: variação da taxa de amostragem; variação de nível de ruído; e variação da intensidade de reverberação. Como medidas acústicas, foram consideradas MFCC e LPC, amplamente utilizadas em tarefas de processamento de voz, como reconhecimento de fala e de locutor [10], [11], reconhecimento de emoções [12], e avaliação de distúrbios da voz [13], [14]. Para fins comparativos, foram empregadas as medidas $F0$, *Jitter* e *Shimmer*, consideradas tradicionais na avaliação de distúrbios vocais [15].

O restante do texto está organizado da seguinte forma: Na Seção II é apresentada a formulação do problema, juntamente com a descrição das características utilizadas. Na Seção III, são apresentados os detalhes do cenário experimental. Na Seção IV são apresentados os resultados obtidos nos experimentos realizados e, na Seção V, é apresentada a conclusão.

## II. Formulação do problema

A análise acústica é um método auxiliar na prática clínica. Com ela, é possível obter mais precisão no que diz respeito à triagem, diagnóstico e tratamento de distúrbios vocais [16]. Tradicionalmente, a voz do paciente é coletada presencialmente, em um ambiente acusticamente controlado. Nesse caso, em geral, o paciente está dentro de uma cabine acústica, emitindo o sinal de voz para ser captado por um microfone profissional. Em contrapartida, considerando um contexto de telessaúde, o ambiente, além de poder ser maior que uma cabine de estúdio, está sujeito a distorções acústicas. Ademais, a captação sonora é realizada por um dispositivo acessível ao paciente, como um *smartphone*.

Entre os fatores que podem influenciar na presença de distorções no sinal de voz captado, estão: a taxa de amostragem do dispositivo, o ruído e a reverberação do ambiente. A taxa de amostragem, por exemplo, pode variar significativamente entre diferentes dispositivos de gravação, impactando a resolução temporal dos sinais vocais. O ruído ambiente, que pode incluir sons de fundo de diversas fontes, interfere na clareza do sinal de voz, dificultando a extração precisa de características acústicas. A reverberação, resultante de reflexões sonoras em ambientes fechados, pode distorcer o sinal de voz captado, introduzindo artefatos que comprometem a análise.

### A. Características Acústicas

A seguir, é apresentada brevemente a descrição das características acústicas empregadas neste trabalho: MFCC, LPC, $F0$, *Jitter* e *Shimmer*.

*1) MFCC:* A medida está relacionada à percepção do ouvido humano, aproximando computacionalmente a percepção auditiva para uma escala chamada Mel [17].

Para a extração do MFCC, após a etapa de pré-processamento com a segmentação do sinal de voz, as amostras de cada quadro são convertidas para o domínio da frequência através da transformada rápida de Fourier (*fast Fourier transform* – FFT), a partir da qual é calculada a energia. O sinal transformado passa então através de um banco de filtros em escala Mel. O conjunto de coeficientes, $c_j$ é obtido de acordo com [18]:

$$c_j = \sum_{k=1}^{F} (\log\ S_k)\ \cos\left[\frac{\pi j}{F}\left(k - \frac{1}{2}\right)\right], \tag{1}$$

para $j = 1,2,...,D$, em que $D$ é o número de coeficientes, $S_k$ é a energia do $k$-ésimo filtro, e $F$ é a quantidade de filtros na escala Mel.

*2) LPC:* O princípio básico da análise LPC é determinar um conjunto de coeficientes preditores, $\alpha(k)$, diretamente do sinal de fala, para obter uma estimativa adequada das propriedades espectrais dos sinais. [14], [19]. Assim, o trato vocal pode ser modelado com a seguinte função de transferência:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha(k) z^{-k}}, \tag{2}$$

em que $G$ é o fator de ganho, ajustado para controlar a intensidade de excitação e $p$ é a ordem do preditor.

*3) F0:* A frequência fundamental $F0$ refere-se à menor frequência de uma onda sonora e é percebida como a altura ($pitch$) da voz [20]. É a frequência na qual as pregas vocais vibram durante a produção de sons vocais:

$$F0 = \frac{1}{T0}, \tag{3}$$

em que $T0$ é o período de um ciclo de vibração das pregas vocais.

*4) Jitter:* É uma medida da variação de curto prazo da frequência $F0$. Representa a instabilidade na frequência de ciclo a ciclo das pregas vocais [4]. Existem diferentes formas de calcular o *Jitter*, mas uma equação comum é o *Jitter* local percentual, que representa a diferença média absoluta entre dois períodos consecutivos, dividido pelo período médio:

$$Jitter = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100\%, \tag{4}$$

em que $T_i$ é o período do $i$-ésimo ciclo e $N$ é o número total de períodos.

*5) Shimmer:* É uma medida da variação de curto prazo na amplitude da onda sonora. Reflete a instabilidade na intensidade de ciclo a ciclo das pregas vocais [4]. O *Shimmer* local percentual representa a diferença média absoluta entre as amplitudes de dois períodos consecutivos, dividido pela amplitude média:

$$Shimmer = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100\%, \tag{5}$$

em que $A_i$ é a amplitude do $i$-ésimo ciclo.

## III. Cenário Experimental

Nesta seção são apresentados os aspectos metodológicos deste trabalho, como a base de dados utilizada e as suas variações de acordo com distorções acústicas. Ainda são descritas as etapas de extração de características e classificação.

### A. Base de dados

A base empregada neste trabalho é a *Saarbruecken Voice Database* (SVD) [21], desenvolvida na Alemanha e distribuída por meio de um repositório digital aberto[1], a qual possui tanto sinais de voz gravados, quanto sinais de eletroglotografia (EGG) capturados em laboratório. Da base, foram coletados 620 sinais de voz da vogal sustentada /a/, sendo 300 sinais de vozes saudáveis, 63 afetados por edema de Reinke, 21 afetados por carcinoma, 41 afetados por leucoplasia, 45 afetados por pólipos vocais e 150 afetados por paralisia das pregas vocais.

A taxa de amostragem considerada na composição da base foi 50 kHz. Neste trabalho adota-se 44,1 kHz para o conjunto de dados original (por meio de subamostragem), para equiparar-se aos mecanismos de gravação de alta qualidade em sistemas de gravação/comunicação [8]. Além disso, para retirar qualquer influência do tamanho amostral nos resultados, foram retirados aleatoriamente, entre todas as patologias, 20

[1] https://stimmdb.coli.uni-saarland.de/index.php4#target, acesso em 06/2024

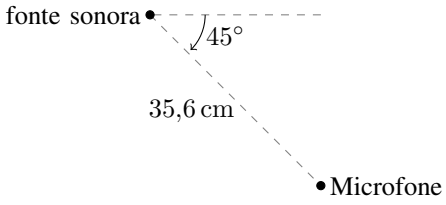


Fig. 1. Posição relativa da fonte sonora e do microfone no ambiente de simulação.

sinais, para que ambas as classes tenham 300 sinais nos experimentos. Em cada classe, saudável e patológica, tem-se 40% das pessoas do gênero masculino e 60% do gênero feminino, com participantes com idade entre 18 e 65 anos.

### *B. Variações do conjunto de dados original*

Com o intuito de avaliar o efeito de distorções acústicas e da influência de gravações realizadas por *smartphones* na classificação de patologias vocais, três versões do conjunto de dados original foram geradas por meio de simulação. Cada versão contemplando a:

i) Variação da taxa de amostragem. Aqui, as gravações do conjunto de dados original foram reamostradas para $16\,\mathrm{kHz}$ e $8\,\mathrm{kHz}$.
ii) Variação do nível de ruído. Nessa versão, foram geradas variações do conjunto de dados original adicionando ruído branco gaussiano (*additive white Gaussian noise* – AWGN), com relação sinal ruído (*signal-to-noise ratio* – SNR) de $0\,\mathrm{dB}$ até $40\,\mathrm{dB}$, com passos de $5\,\mathrm{dB}$. Para evitar o efeito de *clipping*, os áudios gerados que apresentavam amostras fora do intervalo $[-1{,}1]$ foram ajustados por meio de normalização.
iii) Variação de reverberação com o parâmetro RT60[2]. Para isso, foi utilizada a biblioteca *pyroomacoustics*[3] para simular uma sala *shoebox* de $36\,\mathrm{m}^3$, com comprimento, largura e altura de $3\,\mathrm{m}$, $4\,\mathrm{m}$ e $3\,\mathrm{m}$, respectivamente. Como apresentado na Fig. 1, foram colocados nessa sala um microfone, para representar o *smartphone*, e uma fonte sonora representando o locutor. O microfone foi posicionado na frente da fonte sonora, à $35{,}6\,\mathrm{cm}$ e com uma angulação de $45°$ para simular o locutor segurando o dispositivo na mão. A posição do microfone e da fonte sonora, em coordenadas cartesianas, foi $(1{,}56; 2; 1{,}5)\,\mathrm{m}$ e $(1{,}78; 2; 1{,}78)\,\mathrm{m}$. Os microfones foram configurados com padrão de captação omnidirecional e ganho de $0{,}1$. Já a fonte sonora foi configurada com padrão de irradiação subcardioide (em direção ao microfone) e com ganho de $0{,}5$. Por fim, foram geradas variações do conjunto de dados original com valores de RT60 de $0{,}1\,\mathrm{s}$ até $0{,}6\,\mathrm{s}$, com passos de $0{,}1\,\mathrm{s}$.

### *C. Extração de Características*

Para o cenário experimental deste trabalho, foi utilizada a linguagem de programação Python. A segmentação dos sinais de voz ocorreu com um tamanho de quadro de 25 ms, com sobreposição de 10 ms. Para as medidas MFCC e LPC, foram considerados 12 coeficientes, extraídos por meio da biblioteca Librosa[4]. As medidas de $F0$, *Jitter* e *Shimmer* foram obtidas por meio da biblioteca Parselmouth-Praat[5], considerando o intervalo de estimação entre 75 Hz e 600 Hz. Estas configurações foram consideradas para todas as variações da base de dados original SVD. Para cada sinal, foi obtido o valor médio de cada característica ao longo dos quadros, a fim de formar o banco de dados para a classificação.

### *D. Classificação*

A classificação realizada neste trabalho é binária, entre saudável e patológico. Nesta etapa, foi considerada a análise discriminante quadrática (*quadratic discriminant analysis* - QDA), amplamente utilizada na literatura para aplicações tabulares com dados de baixa dimensionalidade na entrada [22]. Para tanto, foi utilizada a bilbioteca *scikit-learn*[6]. Os modelos de QDA consistem em uma generalização dos modelos de análise discriminante linear, permitindo a modelagem de fronteiras de decisão mais flexíveis, baseadas em expressões quadráticas [23]. Para dar mais confiabilidade aos resultados, foi empregado o método *k-fold* de validação cruzada, com $k = 10$, que tem sido comumente utilizado em classificação de distúrbios vocais [16], [24], [25].

Como métricas de desempenho do classificador empregado neste trabalho, foram utilizadas a acurácia, a sensibilidade e a especificidade. A acurácia representa a taxa global de acerto. A sensibilidade está relacionada à capacidade de um classificador em diagnosticar uma doença em um paciente doente, e a especificidade compreende a capacidade de diagnosticar um estado saudável em um paciente saudável [16].

Na prática clínica, de maneira geral, considera-se 70% de acurácia um desempenho de classificação aceitável para o auxílio da avaliação por meio da análise acústica [15].

## IV. Resultados

Nesta seção são apresentados os resultados dos experimentos realizados com as variações da taxa de amostragem, do nível de ruído e de reverberação.

### *A. Variação da Taxa de Amostragem*

Na Tabela I são apresentados os resultados da classificação para cada uma das medidas empregadas, considerando a variação da taxa de amostragem. Os valores de acurácia acima de 70% estão em destaque. Neste contexto, *Shimmer* obteve destaque na acurácia para as três taxas de amostragem. Porém, houve uma discrepância entre os valores de sensibilidade e especificidade, indicando que esta medida não foi capaz de identificar de forma eficiente os casos patológicos. MFCC e LPC obtiveram acurácia acima de 75% com 44,1 kHz e 16 kHz, sendo que para MFCC as métricas de sensibilidade e especificidade se mostraram mais equilibradas entre si quando comparadas àquelas obtidas para LPC. As medidas $F0$ e *Jitter* proporcionaram ao classificador um desempenho abaixo de 70% na acurácia em todos os cenários de amostragem.

[2]O RT60 é a medida de tempo necessário para o nível de pressão sonora decair em 60 dB após o cessar da fonte de áudio.
[3]https://github.com/LCAV/pyroomacoustics, acesso em 06/2024
[4]https://librosa.org/doc/0.10.1/index.html, acesso em 06/2024
[5]https://parselmouth.readthedocs.io/en/stable/, acesso em 06/2024
[6]https://scikit-learn.org/, acesso em 07/2024

TABELA I
DESEMPENHO DO CLASSIFICADOR QDA COM A VARIAÇÃO DA TAXA DE AMOSTRAGEM.

| | Acurácia (%) | | | Sensibilidade (%) | | | Especificidade (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| kHz | 44,1 | 16 | 8 | 44,1 | 16 | 8 | 44,1 | 16 | 8 |
| MFCC | **78,50** | **76,83** | 69,83 | 73,36 | 71,22 | 67,43 | 83,38 | 82,04 | 72,37 |
| LPC | **77,67** | **75,17** | 64,83 | 64,60 | 61,74 | 43,84 | 90,00 | 87,82 | 84,81 |
| $F0$ | 60,83 | 64,67 | 54,67 | 63,98 | 44,01 | 38,16 | 58,86 | 85,31 | 71,29 |
| *Jitter* | 65,83 | 69,83 | 57,17 | 32,84 | 45,86 | 16,16 | 98,27 | 94,32 | 98,17 |
| *Shimmer* | **71,83** | **73,33** | **73,50** | 46,09 | 53,64 | 52,68 | 97,27 | 92,27 | 95,26 |

### B. Variação de SNR

Os resultados de classificação com a variação da SNR são apresentados na Figura 2. No contexto da acurácia (Figura 2a), quando se observa cada medida individualmente, algumas delas parecem ser mais sensíveis ao ruído do que outras. Por exemplo, MFCC e LPC tiveram decaimento progressivo com o aumento do ruído. Outras características, como *Jitter* e *Shimmer* tiveram comportamento mais constante e decaimento da acurácia apenas quando o ruído se tornou mais severo. No entanto, em uma comparação entre as medidas, MFCC e LPC mantiveram acurácia acima de 70% em todos os cenários ruidosos e MFCC continuou tendo o melhor desempenho em quase todos os cenários. Vale notar que *Shimmer* foi a métrica que mais sofreu em termos de acurácia quando o ruído se tornou mais severo.

No contexto da sensibilidade (Figura 2b), apenas a medida MFCC foi a que teve a capacidade de manter a sensibilidade acima de 70%, o que ocorreu apenas em cenários praticamente sem ruído. No entanto, vale notar que apresentou grande variação nos valores, apresentando um ponto de mínimo com SNR de 10 dB. Em relação à especificidade (Figura 2c), todas as medidas, exceto $F0$, apresentaram desempenho acima de 70%. Em particular MFCC, que obteve os maiores valores de acurácia, manteve a especificidade acima de 80% em todos os cenários de SNR.

### C. Variação de Reverberação

Na Figura 3 são apresentados os resultados da classificação realizada a partir da variação do parâmetro de reverberação RT60. No que diz respeito à acurácia (Figura 3a), MFCC e LPC proporcionaram ao classificador um desempenho superior em comparação às outras características, sendo MFCC melhor em praticamente todos os cenários de RT60, com acurácia em torno de 80%. LPC, por outro lado, teve queda de mais de 5 pontos percentuais na acurácia com o aumento da reverberação. Das outras medidas, *Shimmer* chegou a ter desempenho acima de 70% com menor RT60, porém teve queda na acurácia quando a reverberação aumentou. As medidas de $F0$ e *Jitter* tiveram desempenho abaixo de 70% em praticamente todos os cenários de RT60.

Quanto à sensibilidade (Figura 3b), MFCC foi a única medida que teve a capacidade de manter o desempenho acima de 70%, com comportamento similar mesmo com a variação do RT60. Entre as demais características, LPC e $F0$ apresentaram comportamento semelhante em alguns casos de reverberação, ficando ambas com desempenho variando entre 60% e 70% de sensibilidade. No contexto da especificidade (Figura 3c), ocorreu um comportamento similar ao que aconteceu com a variação da SNR: à exceção de $F0$, todas as medidas apresentaram especificidade acima de 70%. Vale notar que as medidas

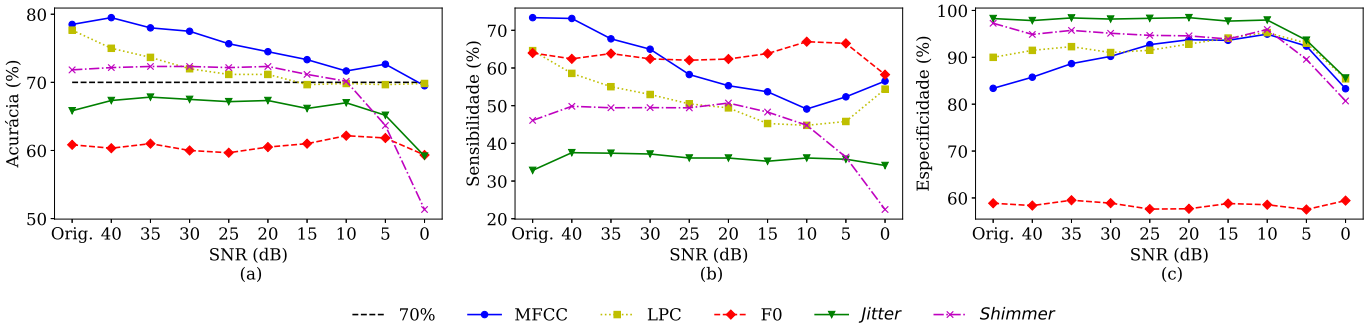


Fig. 2. Comparação dos resultados de classificação do banco de dados original (Orig.) com suas variações em relação à SNR: (a) Acurácia; (b) Sensibilidade; (c) Especificidade.

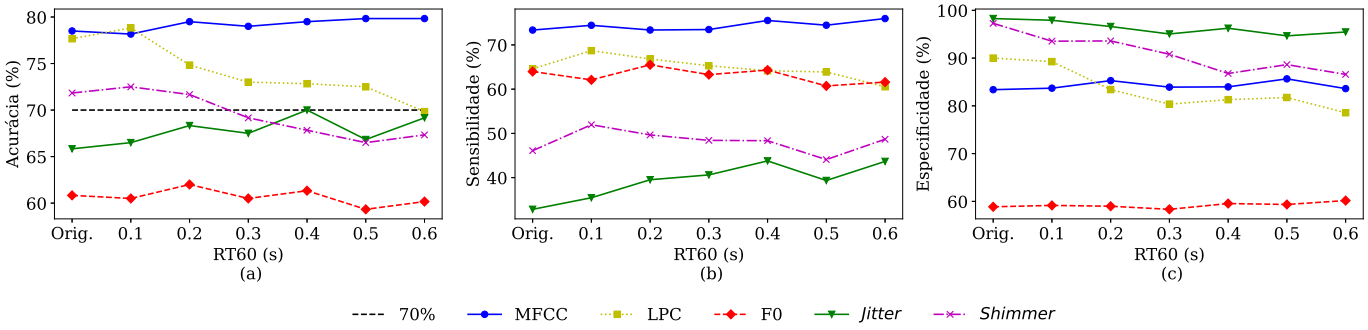


Fig. 3. Comparação dos resultados de classificação do banco de dados original (Orig.) com suas variações em relação à reverberação (RT60): (a) Acurácia; (b) Sensibilidade; (c) Especificidade.

*Jitter* e $F0$ apresentaram uma contraposição em relação aos desempenhos em termos de sensibilidade e especificidade, o que indica a tendência a prover classificações positivas ou negativas em maior quantidade, fato que é refletido no desempenho mais baixo em termos de acurácia dessas medidas.

### *D. Discussão*

A busca por características acústicas confiáveis no contexto de patologias laríngeas é uma tarefa desafiadora, pois depende de fatores como o tipo e a gravidade da lesão e a severidade dos seus efeitos [24], [25]. Com o emergente uso da telessaúde na prática clínica, as distorções acústicas tornam-se outros fatores desafiadores.

Traçando um comparativo entre as medidas empregadas neste trabalho, os resultados mostraram que MFCC pode ser uma escolha razoável caso a necessidade seja trabalhar com uma taxa de amostragem mais baixa que a original. Com 16 kHz, esta medida proporcionou ao classificador desempenho superior a 70% em termos de acurácia, sensibilidade e especificidade. Em um contexto de ruído AWGN, MFCC se sobressaiu em relação às outras medidas, porém apresentou decaimento da sensibilidade para SNRs mais baixas. Em relação à reverberação, MFCC também superou as demais medidas, sendo que os seus valores de acurácia, sensibilidade e especificidade permaneceram quase constantes para todos os níveis de RT60 considerados nos experimentos.

De maneira geral, a análise realizada neste trabalho demonstrou que, embora seja possível extrair informações relevantes da voz sob condições adversas, a variabilidade introduzida pelas distorções acústicas requer a aplicação de técnicas robustas de compensação para assegurar a confiabilidade das características extraídas. Uma abordagem para isso pode ser a combinação de características, como realizado em [26], em que se utilizam medidas como *Jitter* e *Shimmer* em sinais sem distorções. Ou ainda, pode se fazer uso de técnicas de aprendizado de máquina com redes neurais, como em [27], em que foi empregada a medida MFCC em sinais sem distorções.

## V. Conclusões

Este trabalho realizou uma análise das variações de distorções acústicas e de seu efeito na classificação de patologias laríngeas. Os resultados da classificação, realizada com QDA, indicaram que as características acústicas apresentam variação de desempenho na presença de distorções. Dentre elas, MFCC se mostrou mais robusta em termos de acurácia, sensibilidade e especificidade. Estes achados ressaltam a importância de desenvolver métodos aprimorados para a mitigação das influências ambientais, garantindo que a análise acústica permaneça uma ferramenta viável e precisa para a telessaúde. Em trabalhos futuros, pretende-se: realizar a combinação de características acústicas na etapa de classificação; investigar outras medidas além das empregadas neste trabalho; usar outros classificadores; e experimentar o uso simulado de dois microfones do *smartphone* para captação do sinal de voz.

## Agradecimentos

Os resultados apresentados neste artigo foram desenvolvidos como parte de um projeto do SiDi, financiado pela Samsung Eletrônica da Amazônia Ltda., com o apoio da Lei Federal de Informática no. 8248/91.

## Referências

[1] A. Chern et. al., "A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom," *IEEE Access*, vol. 5, pp. 10 339–10 351, 2017.

[2] G. Cantarella et al., "The challenge of virtual voice therapy during the covid-19 pandemic," *Journal of Voice*, vol. 35, no. 3, pp. 336–337, 2021.

[3] S. S. Wang et al., "Continuous speech for improved learning pathological voice disorders," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 3, pp. 25–33, 2022.

[4] M. S. Behlau, "Voz o livro do especialista. Revinter," 2001.

[5] N. V. Mallipeddi, A. Mehrotra, and J. H. Van Stan, "Telepractice in the treatment of speech and voice disorders: What could the future look like?" *Perspectives of the ASHA Special Interest Groups*, vol. 8, no. 2, pp. 418–423, 2023.

[6] B. LaBarge et al., "Comparison of voice therapy outcomes: clinic vs telehealth," *Journal of Voice*, 2023.

[7] J. Lebacq et al., "Maximal ambient noise levels and type of voice material required for valid use of smartphones in clinical voice research," *Journal of voice*, vol. 31, no. 5, pp. 550–556, 2017.

[8] S. Jannetts et al., "Assessing voice health using smartphones," *Int. J. of Lang. & Comm. disorders*, vol. 54, no. 2, pp. 292–305, 2019.

[9] M. G. Di Cesare et al., "Assessment of voice disorders using machine learning and vocal analysis of voice samples recorded through smartphones," *BioMedInformatics*, vol. 4, no. 1, pp. 549–565, 2024.

[10] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[11] P. Bansal, S. A. Imam, and R. Bharti, "Speaker recognition using MFCC, shifted MFCC with vector quantization and fuzzy," in *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*. IEEE, 2015, pp. 41–44.

[12] M. S. Fahad et al, "DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features," *Circuits, Systems, and Signal Processing*, vol. 40, pp. 466–489, 2021.

[13] S. Tirronen, S. R. Kadiri, and P. Alku, "The effect of the MFCC frame length in automatic voice pathology detection," *Journal of Voice*, 2022.

[14] J. R. Orozco-Arroyave et al., "Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases," *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1820–1828, 2015.

[15] L. W. Lopes et al. "Acurácia das medidas acústicas tradicionais e formânticas na avaliação da qualidade vocal," in *CoDAS*, vol. 30. SciELO Brasil, 2018, p. e20170282.

[16] W. C. de Almeida Costa, "Análise dinâmica não linear de sinais de voz para detecção de patologias laríngeas," Ph.D. dissertation, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Campina Grande, 2012.

[17] D. O'shaughnessy, *Speech communication: human and machine*. Universities press, 1987.

[18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Sig. Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[19] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-Hall Signal Processing Letters, 1978.

[20] G. Fant, *Acoustic theory of speech production: with calculations based on X-Ray studies of russian articulations*. Walter de Gruyter, 1971.

[21] B. Woldert-Jokisz, "Saarbruecken voice database," 2007.

[22] R.S.-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer, 2009.

[24] V. J. D. Vieira, "Avaliação de distúrbios da voz por meio de análise de quantificação de recorrência," Master's thesis, Programa de Pós-Graduação em Engenharia Elétrica, Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, 2014.

[25] L. W. Lopes et. al., "Accuracy of acoustic analysis measurements in the evaluation of patients with different laryngeal diagnoses," *Journal of Voice*, vol. 31, no. 3, pp. 382–e15, 2017.

[26] L. Lopes, V. Vieira, and M. Behlau, "Performance of different acoustic measures to discriminate individuals with and without voice disorders," *Journal of Voice*, vol. 36, no. 4, pp. 487–498, 2022.

[27] X. Xie, H. Cai, C. Li, Y. Wu, and F. Ding, "A voice disease detection method based on MFCCs and shallow CNN," *Journal of Voice*, 2023.