Avaliação da soprosidade vocal em amostras codificadas pelo codec IETF Opus

João Pedro Hallack Sansão, Leonardo Carneiro de Araújo, Hani Camille Yehia e Maurílio Nunes Vieira

***Resumo*— O comportamento de correlatos acústicos à soprosidade vocal foi investigado em diferentes níveis de compressão. O codec utilizado foi o IETF Opus com bitrates entre 16 e 256 kbps. As medidas escolhidas foram: *Smoothed cepstral peak prominence* (CPPS), *spectral flatness of the residue signal* (SFRS), *pitch amplitude* (PA) e *spectrographic signal-to-noise-ratio* (S^2NR). A investigação foi conduzida em duas etapas: inicialmente usando amostras de voz sintética que tinham como referência a relação sinal-ruído controlada e posteriormente amostras de voz natural classificadas perceptivamente em soprosidade. As medidas acústicas mantiveram alta correlação com as respectivas referências: SNR para voz sintética e soprosidade para voz natural. O erro quadrático médio estabilizou-se na maior parte das medidas até a taxa de 128 kbps. Considerando os resultados, o Opus não comprometeu a avaliação da soprosidade e sua adoção irá permitir economia em armazenamento e largura de banda na transmissão das amostras.**

***Palavras-Chave*— Avaliação de voz, qualidade de voz, voz disfônica, correlatos acústicos, diagnóstico remoto, telemedicina, soprosidade, relação sinal-ruído, codec livre, IETF Opus, RFC 6716**

***Abstract*— The behavior of acoustic correlates of breathy vocal quality was investigated under different levels of compression. The IETF Opus codec was employed with varying bitrates from 16 to 256 kbps. The tested measures were: smoothed cepstral peak prominence (CPPS), spectral flatness of the residue signal (SFRS), pitch amplitude (PA) and spectrographic signal-to-noise-ratio (S^2NR). The investigation was conducted in two parts: using synthetic voice samples with controlled signal-to-noise ratio and natural voice perceptually qualified in breathiness levels. Acoustic measures showed high correlation levels with their references: SNR for synthetic voice and breathiness for natural voice. In most measures, root-mean-square error is stable before attaining 128 kbps. Considering the results, the Opus codec did not compromise the breathiness evaluation and its adoption will allow savings in storage and bandwidth.**

***Keywords*— Voice evalutaion, voice quality, dysphonic voice, acoustic correlates, remote assessment, telemedicine, breathiness, signal-to-noise ratio, open-source codec, IETF Opus, RFC 6716**

I. Introdução

A análise acústica da voz visa descrever parâmetros objetivos para discriminar falantes disfônicos dos não-disfônicos, correlacionar um determinado parâmetro à avaliação perceptiva da qualidade de voz e permitir o tratamento longitudinal dos pacientes[1].

Um dos parâmetros perceptivos que a literatura correlaciona fortemente a medidas acústicas é a soprosidade [2]. Este fenômeno está relacionado ao ruído glótico, causado pela passagem do fluxo turbulento de ar devido ao fechamento incompleto da glote durante o ciclo fonatório [3]. Diversas medidas estão disponíveis para o estudo da soprosidade [1], [2], [4].

A análise acústica da voz depende basicamente do registro das elocuções em meio digital para posterior processamento. É um método não-invasivo e de baixo custo devido aos recentes avanços na computação pessoal (incluindo dispositivos móveis com poder computacional), ferramentas para registro e transmissão de multimídia. A tecnologia abriu a possibilidade de realização de triagens, monitoramento e avaliação da qualidade de voz de forma remota.

Com freqüência, devido às limitações de largura de banda para transmissão e de espaço de armazenamento, se faz necessário comprimir os arquivos. Diversos codecs estão disponíveis para esta tarefa, com variadas características. Como, em geral, esta compressão é realizada com perdas, deseja-se avaliar o comportamento das medidas acústicas com a diminuição da taxa de transmissão usada na codificação.

Em um estudo preliminar [5], a relação de correlatos acústicos com variação da taxa de compressão dos codecs livres SPEEX[6], CELT [7] e Ogg Vorbis [8] foi estudada. Como cada um deles trabalhava em uma faixa de bitrate diferente, não foi possível determinar um único codec que teria melhor comportamento em uma maior faixa possível.

Em 2013, atendendo aos requisitos da RFC 6716 da IETF, foi lançado o codec Opus [9], [10], que visa suplantar através de uma única implementação, o SPEEX, CELT e o Ogg Vorbis. Com esta unificação, é possível com um único codec, trabalhar em uma maior faixa de bitrate.

O objetivo deste trabalho é avaliar o comportamento de algumas medidas acústicas usando o IETF Opus sob variadas taxas de compressão.

A. Correlatos acústicos da soprosidade

Em busca de uma sistematização para comparar as medidas acústicas com esta percepção geral da qualidade de voz, um estudo sistemático foi realizado em [4]. Como os dados de diferentes estudos envolviam diferentes metodologias e critérios perceptivos diferentes (diferentes protocolos, tamanho e tipo de escala perceptiva, qualidade avaliada, etc), os autores do referido estudo recorreram à meta-análise, que por sua vez, aponta que para a vogal sustentada, as melhores medidas acústicas são: *Smoothed cepstral peak prominence* (CPPS), *spectral flatness of the residue signal* (SFRS), *pitch amplitude* (PA), que foram utilizadas neste trabalho.

João Pedro Hallack Sansão e Leonardo Carneiro de Araújo¸ Campus Alto Paraopeba, Universidade Federal de São João del-Rei, Ouro Branco-MG, Brasil, e-mail para correspondência: joao@ufsj.edu.br. Hani Camille Yehia e Maurílio Nunes Vieira, Departamento de Engenharia Eletrônica, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil.

Além das medidas sugeridas pela meta-análise, a relação sinal-ruído espectrográfica S^2NR [1] também foi incluída nas medidas a serem estudadas. A escolha da vogal sustentada como forma de elocução é justificada pela maior estacionaridade do sinal em relação à fala corrente.

As medidas escolhidas têm alta correlação com a dimensão perceptiva da soprosidade:

1) Smoothed cepstral peak prominence (CPPS): No estudo meta-analítico [4], 36 medidas para vogal sustentada e 3 para fala contínua foram comparadas. A medida *Smoothed cepstral peak prominence* (CPPS[2]) emergiu como a medida mais robusta do estudo. CPPS não depende na detecção precisa de picos, é adequada para vogais sustentadas e fala contínua e é facilmente implementada. É calculada como a diferença suavizada na amplitude entre o logaritmo do pico do cepstrum e "ruído de fundo", representado através de uma regressão linear.

2) Pitch amplitude (PA): Neste método [11], calcula-se os coeficientes de predição linear do sinal. Com estes coeficientes, efetua-se a filtragem inversa do sinal de voz, obtendo o resíduo do sinal.

Calcula-se então a função de autocorrelação normalizada do resíduo (operando em janelas de 100 ms). Buscam-se, então, os picos da autocorrelação. O valor da função de autocorrelação neste pico é valor da medida na janela, designada *Pitch amplitude*.

Para um sinal perfeitamente periódico, a PA vale 10,0 (na normalização empregada por [11], também utilizada na implementação deste trabalho).

3) Spectral flatness of the residue signal (SFRS): A SFRS [12] por sua vez é baseada no espectro do resíduo da predição linear. Sua definição, dada em decibéis, é de SFRS $= 10\log_{10}(\epsilon)$, onde $0 \leq \epsilon \leq 1$, e ϵ é a razão da média geométrica com a média da energia espectral. Idealmente, $\epsilon = 1$, para o caso de um espectro perfeitamente plano de ruído não correlacionado. Desta forma, a medida tende 0 dB com perturbação crescente e tende a $-\infty$ com periodicidade crescente.

4) Relação sinal-ruído espectrográfica (S^2NR): S^2NR estima a relação sinal-ruído a partir da imagem do espectrograma de banda-estreita. Utilizando técnicas de processamento de sinais bidimensionais utilizados para identificação de impressões digitais[1], esta técnica tem como objetivo principal a redução da dependência à detecção precisa da freqüência fundamental e das sensibilidades às aperidiocidades do sinal.

Seu funcionamento é baseado na orientação das linhas harmônicas da imagem do espectrograma. A orientação sintoniza as etapas do algoritmo que geram máscaras binárias que definem as regiões de sinal e ruído, determinando assim a relação sinal-ruído espectrográfica.

Diversos benchmarks mostram que o método é mais robusto às pertubações em amplitude e freqüência que os citados anteriormente [1].

B. IETF Opus

O codec Opus [9], [10] é um codec de áudio interativo em tempo-real para atender os requisitos descritos na RFC 6716. É composto por uma camada baseada em predição linear [12] e uma camada baseada na MDCT (transformada discreta de cossenos modificada[13]).

A idéia básica de usar as duas camadas é que na fala, técnicas baseadas em predição linear (como CELP) codificam baixas freqüências de forma mais eficiente do que métodos baseados em transformadas no domínio da freqüência, como a MDCT, enquanto quando se codifica música e fala em freqüências mais altas, a situação se reverte [9].

Desta forma, um codec com ambas camadas pode operar em uma faixa mais abrangente do que um codec simples e pode exibir melhor qualidade combinando ambas as características.

O codec Opus é livre (grátis e código aberto) e disponibilizado sob licença BSD. As patentes que cobrem seu funcionamento são livres de *royalties*.

Com as características expostas, ele suplanta os codecs livres para codificação de fala como SPEEX [6] e CELT [7], e o codec Ogg Vorbis [8] para codificação de música.

II. Estímulos e experimentos

As avaliações foram feitas em duas etapas: sinais de voz sintética e sinais de voz natural.

A implementação das medidas instrumentais utilizadas neste estudo está disponível sob licença livre [14], [15].

A. Voz sintética

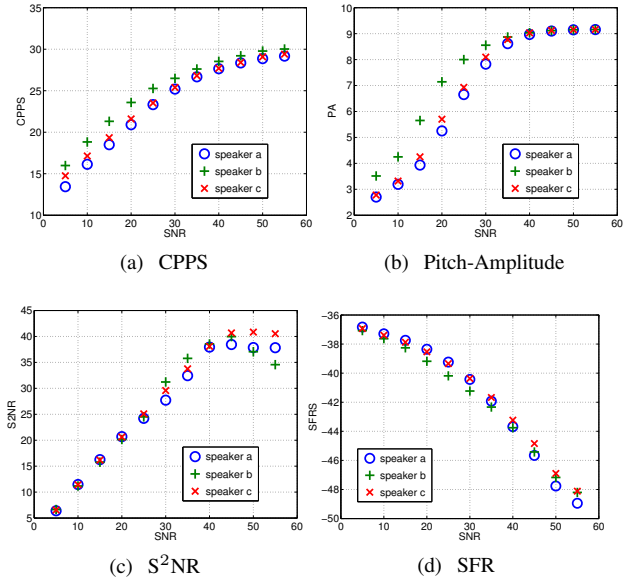
Foram utilizados sinais de voz sintetizados com relação sinal-ruído conhecida para as avaliações iniciais. A vogal sintetizada /a/ (amostragem de 22050 Hz, 16 bits por amostra) foi criada convoluindo pulsos glotais com a resposta ao impulso do trato vocal, a radiação labial sendo modelada por um diferenciador de primeira ordem no fluxo bucal. O modelo de fluxo glótico [16] utilizado nos experimentos tem fator de inclinação (*skew*) de $K = 0{,}65$. A resposta ao impulso do trato vocal foi estimada utilizando análise com codificação preditiva linear (LPC) com os parâmetros de 95% de pré-ênfase, 22^a ordem, método da covariância [12] da vogal /a/ de três falantes distintos.

O ruído foi adicionado ao sinal $s(n)$ para produzir sinais com relação sinal-ruído conhecida a priori, independente da filtragem do trato vocal. Um sinal ruidoso $x(n)$ foi criado adicionando uma seqüência aleatório $e(n)$ com média nula, uniformemente distribuída ao ciclos de $s(n)$. Este tipo de síntese falha em termos de qualidade perceptiva, mas permite controle fino das propriedades de interesse nos estímulos.

Sinais de referência com relação sinal-ruído pré-definida foram gerados, partido de $\text{SNR}_{\text{ref}} = 5\text{dB}$ até 55dB, com passo de 5dB, para os três falantes (três tratos distintos, com a vogal /a/). A freqüência fundamental do pulso glótico é de $f_o = 120\,\text{Hz}$. A Figura 1 mostra a relação das medidas acústicas com relação sinal-ruído de referência, destacando a alta correlação das medidas com o parâmetro de referência.

Os sinais de referência, em seqüência, foram codificados usando o codec Opus, variando as taxas nominais de bitrate de 16 kbps até 256 kbps com passos de 16 kbps.

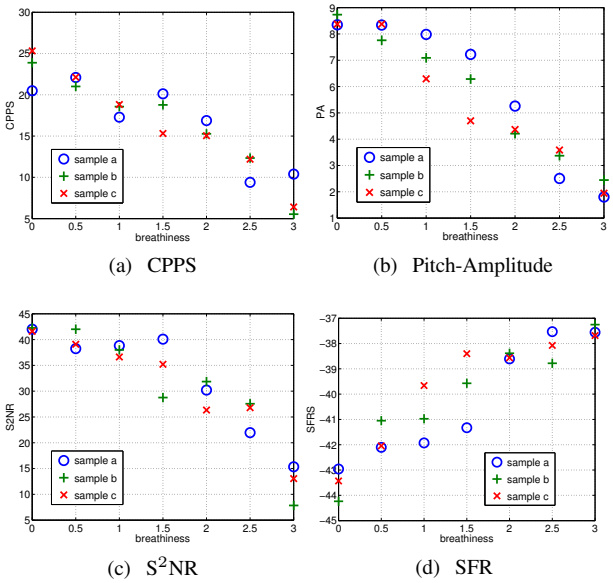
Para cada uma das amostras geradas foram calculadas as quatro medidas acústicas. Dos resultados das medidas



(a) CPPS (b) Pitch-Amplitude

(c) S^2NR (d) SFR

Fig. 1. Medidas acústicas versus SNR de referência em voz sintética, vogal /a/ para três tratos distintos



(a) CPPS (b) Pitch-Amplitude

(c) S^2NR (d) SFR

Fig. 2. Medidas acústicas versus avaliação perceptiva da soprosidade, vogal /a/ para três falantes distintos em cada nível

calculou-se então, tomando como referência os valores obtidos sem compressão: o erro quadrático médio absoluto para cada medida, sob um determinado bitrate, o erro quadrático médio relativo e o índice de correlação de Pearson da medida acústica com a relação sinal ruído de referência, para cada bitrate.

B. Voz natural

Para estudar a relação entre os valores das medidas citadas e avaliação perceptiva da soprosidade, serviu-se de um banco de dados formado pela a vogal sustentada /a/ de 21 indíviduos adultos. As gravações, com três amostras por nível de severidade foram selecionadas e foram perceptualmente avaliados em uma escala de 7 pontos (ausente, pequeno, moderado, extremo, com níveis intermediários). A escala perceptiva de soprosidade vai de 0 (ausente) a 3 (grau severo), passo de 0,5. As gravações selecionadas (22050 amostras por segundo, 16 bits por amostra) tinham duração superior a 3 segundos, predominatemente soprosas e tinham o nível de soprosidade praticamente estável ao longo da elocução [1]. Em falantes disfônicos, é comum a ocorrência de instabilidades fonatórias, principalmente nas amostras de soprosidade extrema.

Para cada gravação, inicialmente, foram calculados os valores médios de S^2NR, CPPS, PA, SFR e comparados com os respectivas avaliações de soprosidade.

A Figura 2 mostra a relação das medidas acústicas com a avaliação perceptiva da soprosidade, exibindo alta correlação entre o grau de soprosidade e o valor das medidas acústicas.

Da mesma forma que as amostras de voz sintética, as amostras de voz natural em seqüência foram codificadas usando o codec Opus, variando as taxas nominais de bitrate de 16 kbps até 256 kbps com passo de 16 kbps.

Para cada uma das amostras geradas foram calculadas as quatro medidas acústicas. Dos resultados das medidas, calculou-se então, tomando como referência os valores obtidos sem compressão: o erro quadrático médio absoluto para cada medida, sob um determinado bitrate, o erro quadrático médio relativo e o índice de correlação de Pearson da medida acústica com a relação ao grau de soprosidade, para cada bitrate.

III. Resultados e discussão

A. Avaliação com voz sintética

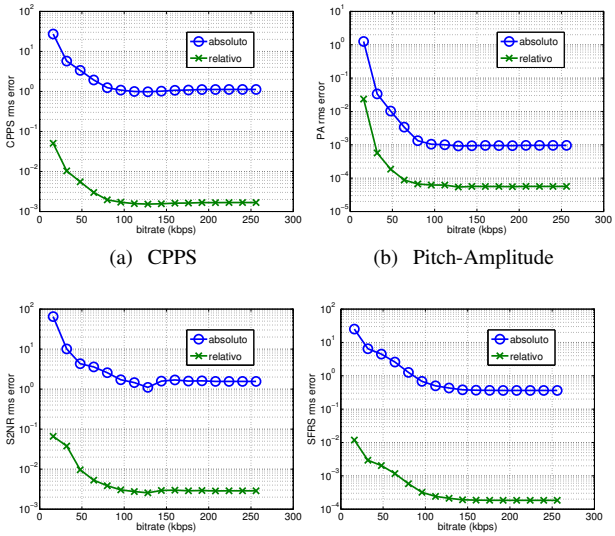
Após o processamento do lote, num total de 561 amostras (11 valores de SNR_{ref}, 3 tratos vocais distintos, 16 valores de bitrate), foram traçados os gráficos da Figura 3. Cada marcador no gráfico indica o erro quadrático médio (absoluto ou relativo) incluindo todas as amostras com a mesma taxa de compressão.

Nas quatro medidas, notou-se que o erro estabiliza-se com o aumento do bitrate. No caso da medidas S^2NR e SFR, o erro quadrático médio praticamente atinge o valor final com taxas em torno de 128 kbps. Na medidas de CPPS, o erro é estabilizado na faixa de 96 kbps e na pitch-amplitude é estabilizado na faixa de 80 kbps.

Este experimento indica que um bitrate superior ao valor de 128 kbps não implica em diminuição do erro de estimação das medidas acústicas, apenas no aumento do tamanho do arquivo. Para referência, um dos arquivos originais, no formato WAV mono ocupa 130 kB de armazenamento. Codificado com bitrates de 64 kbps, 128 kbps e 256 kbps, ocupa 22 kB, 47 kB e 91 kB respectivamente.

Do mesmo conjunto de dados, extraiu-se a correlação do valor obtido pelas medidas e a relação sinal-ruído de referência, para cada bitrate. O resultado está apresentado na Figura 4. Observou-se que mesmo em situações com bitrate baixo, o índice de correlação foi preservado alto e apresentou pequena flutuação com o aumento de bitrate. Os valores de correlação se estabilizaram na faixa entre 0,9-0,95 na CPPS e na Pitch-amplitude, e na faixa entre 0,95-1,0 na SFR e S^2NR. A SFR apresenta correlação negativa com a relação sinal-ruído. As outras medidas apresentam correlação positiva.

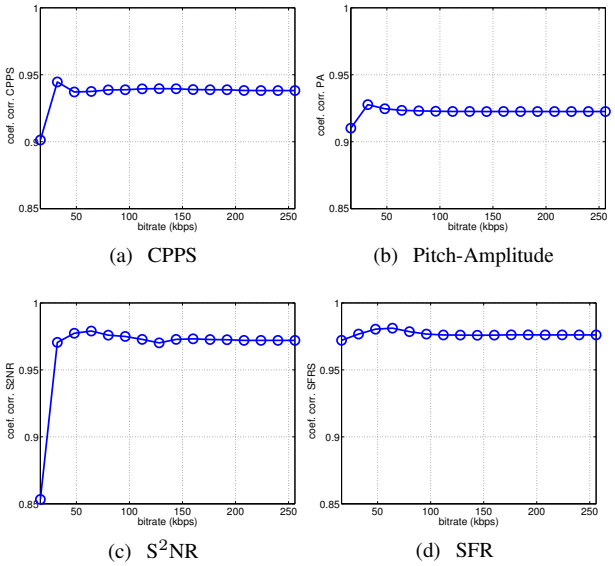
Este resultado indica que mesmo com amostras deterioradas pela compressão, as características de alta correlação com a relação sinal-ruído das medidas são mantidas.



(a) CPPS (b) Pitch-Amplitude

(c) S^2NR (d) SFR

Fig. 3. Avaliação do erro quadrático médio com voz sintética



(a) CPPS (b) Pitch-Amplitude

(c) S^2NR (d) SFR

Fig. 4. Correlação da medida com SNR de referência variando o bitrate com voz sintética

B. *Avaliação com voz natural*

Os resultados com voz sintética indicam a possibilidade do uso das medidas acústicas em amostras de voz real comprimidas.

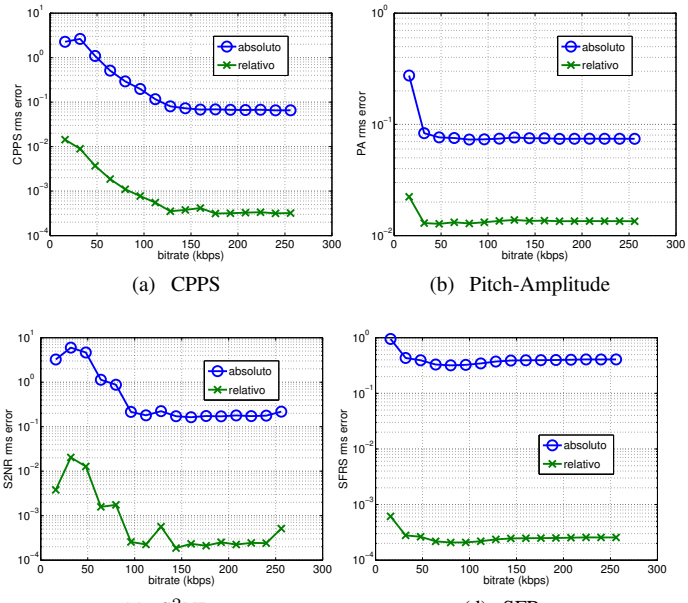
A Figura 5 mostra a evolução do erro quadrático médio com o aumento do bitrate. Para a CPPS e S^2NR, a estabilização do erro ocorreu em taxas superiores a 128 kbps. Notou-se também que o erro quadrático médio foi inferior ao valor obtido em voz sintética. O erro da SFR se estabilizou em baixas taxas, cerca de 64 kbps, com erro na ordem de grandeza do caso sintético. No caso da pitch-amplitude, apesar do erro ter se estabilizado em baixas taxas, notou-se um aumento de 2 ordens de grandeza no erro relativo em relação ao caso sintético.

De forma análoga ao caso anterior, onde calculou-se a correlação da SNR_{ref} com as medidas acústicas para cada taxa de compressão, foi calculada a correlação das medidas acústicas com a avaliação perceptiva de soprosidade.

A evolução do coeficiente de correlação com o bitrate está mostrada na Figura 6. A linha verde continua indica o índice de correlação para o caso sem compressão. Neste caso, notou-se estabilização do valor de índice de correlação na CPPS, Pitch-Amplitude e SFR com baixas taxas. A S^2NR necessitou atingir taxas superiores a 100 kbps para estabilizar no valor de correlação próximo ao ideal. Em todos os métodos, para taxas superiores a 64 kbps, o valor da correlação foi superior a 0,9.

As análises anteriores foram feitas com o módulo do coeficiente de correlação. Considerando o sinal do coeficiente, a SFR apresentou correlação positiva com o grau de soprosidade. As outras medidas apresentaram correlação negativa.

Os altos índices de correlação indicam que as medidas, mesmo com amostras comprimidas, mantiveram-se coerentes com a avaliação perceptiva.



(a) CPPS (b) Pitch-Amplitude

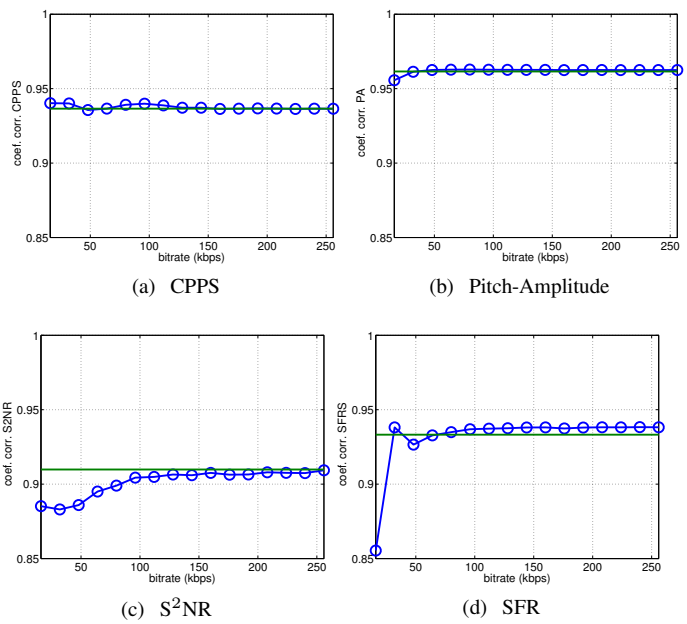
(c) S^2NR (d) SFR

Fig. 5. Avaliação do erro quadrático médio em voz natural

IV. Conclusões

Este trabalho apresentou o comportamento de medidas acústicas correlatas da soprosidade, em voz sintética e voz natural classificada perceptivamente, usando amostras comprimidas em diversas taxas utilizando o codec IETF Opus.

Notou-se que em ambos os casos, o erro quadrático médio é estabilizado em bitrates na faixa de 128 kbps, principalmente para as medidas consideradas mais robustas como a CPPS e



(a) CPPS (b) Pitch-Amplitude

(c) S^2NR (d) SFR

Fig. 6. Correlação com a avaliação perceptiva em voz natural

S^2NR. Este fato demonstra que os valores medidos na amostra original são próximos dos valores na amostra comprimida. Aumentar a taxa de transmissão além desta faixa não reduz o erro quadrático médio.

Pode-se concluir que a utilização do codec Opus não afeta a correlação das medidas com a dimensão perceptiva da soprosidade, mesmo com taxas baixas.

Os resultados indicam que o codec Opus possui características que não interferem significativamente na avaliação da voz soprosa, permitindo economia de espaço de armazenamento e largura de banda na transmissão das amostras.

AGRADECIMENTOS

Agradecimentos à FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) pelo auxílio financeiro.

REFERÊNCIAS

[1] M. N. Vieira, J. P. H. Sansão, and H. C. Yehia, "Measurement of signal-to-noise ratio in dysphonic voices by image processing of spectrograms," *Speech Communication*, vol. 61, pp. 17–32, 2014.

[2] J. Hillenbrand and R. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of speech and hearing research*, vol. 39, no. 2, p. 311, 1996.

[3] J. Laver, S. Wirz, J. Mackenzie, and S. Hiller, "A perceptual protocol for the analysis of vocal profiles," *Work in Progress*, vol. 14, pp. 139–155, 1981.

[4] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," *The Journal of the Acoustical Society of America*, vol. 126, p. 2619, 2009.

[5] J. P. H. Sansão, H. C. Yehia, and M. N. Vieira, "Evaluation of breathiness acoustic correlates under different compression levels," *10th International Conference Advances in Quantitative Laryngology*, p. 109, 2013.

[6] J.-M. Valin, "The speex codec manual version 1.2 beta 3," *Xiph. org Foundation*, 2007.

[7] J.-M. Valin, T. B. Terriberry, C. Montgomery, and G. Maxwell, "A high-quality speech and audio codec with less than 10-ms delay," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 1, pp. 58–67, 2010.

[8] J.-x. YAN, Z.-w. DONG, *et al.*, "Ogg vorbis digital audio coding techniques [j]," *Audio Engineering*, vol. 9, p. 001, 2003.

[9] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the opus codec," in *Audio Engineering Society Convention 135*, Audio Engineering Society, 2013.

[10] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," *IETF, September*, 2012.

[11] R. A. Prosek, A. A. Montgomery, B. E. Walden, and D. B. Hawkins, "An evaluation of residue features as correlates of voice disorders," *Journal of communication disorders*, vol. 20, no. 2, pp. 105–117, 1987.

[12] J. E. Markel and A. H. Gray, *Linear prediction of speech*. Springer-Verlag New York, Inc., 1982.

[13] K. Rao and P. Yip, *Discrete cosine transform*. Academic Press, 1990.

[14] J. P. H. Sansão, "Instrumental measures for voice evalutation." `https://github.com/jsansao/insmes_tools`, 2011.

[15] J. P. H. Sansão, "Spectrographic signal noise ratio measurement." `https://github.com/jsansao/s2nr`, 2011.

[16] G. Fant, "Glottal source and excitation analysis," *Speech Transmission Laboratory - Quaterly Progress and Status Report*, vol. 1, pp. 85–107, 1979.