Análise da perturbação de período de vocalizações humanas e sua aplicação para o aprimoramento da sintese de voz

Giovanna Castilho Davatz Lopes[1]
giovannadavatz@alumni.usp.br
Arlindo Neto Montagnoli[2]
arlindo@ufscar.br
Programa de Pós-Graduação Interunidades em Bioengenharia EESC/IQSC/FMRP da Universidade de São Paulo - USP; e Programa de Pós-Graduação em Engenharia Elétrica – PGEE – da Universidade Federal de São Carlos – UFSCar, São Carlos/SP, Brasil.

INTRODUÇÃO. O movimento das pregas vocais possui modificações que fazem variar o tempo de duração de cada ciclo e consequentemente, o comprimento de onda durante a fonação. Esta perturbação tem sido utilizada como base para cálculo do jitter, aspecto quantitativo desse parâmetro (FARRÚS; HERNANDO, 2009). Tem relevância perceptual, visto que sua remoção de sinais torna o som resultante robótico, distante do natural humano (KERSTA; BRICKER; DAVID, 1960). Pesquisadores têm descrito métodos para gerá-lo: alternando comprimentos de onda, utilizando ruído branco ou flutuações sinusoidais, métodos estocásticos, dentre outros (AOKI; IFUKUBE, 1996; CATALDO; SOIZE, 2017; TEIXEIRA; FERNANDES, 2014). Diante dessa temática, os presentes autores utilizaram uma estratégia diferente, expressa pelos objetivos do estudo – avaliar a perturbação de período presente na voz humana, selecionar modelos matemáticos para simular essa perturbação, incorporar esses modelos na síntese de vogais e comparar a qualidade sonora por meio de julgamento perceptivo-auditivo.

MATERIAIS E MÉTODOS. Estudo experimental e prospectivo aprovado pelo Comitê de Ética em Pesquisa da UFSCar, com protocolo número 2.639.600. Para sua realização foi feita avaliação de sinais de emissões humanas – vogal /a/ sustentada - sendo obtida a perturbação de período por meio de correlação parcial. Este procedimento é feito diretamente no sinal vocal, sendo que uma amostra contendo um único período fundamental é comparada ao sinal inteiro. A correlação cruzada determina picos, permitindo a identificação da variação do período em todo o sinal, o qual foi transformado em traçado gráfico para observação. A partir de sua análise selecionou-se modelos matemáticos para simulá-lo.

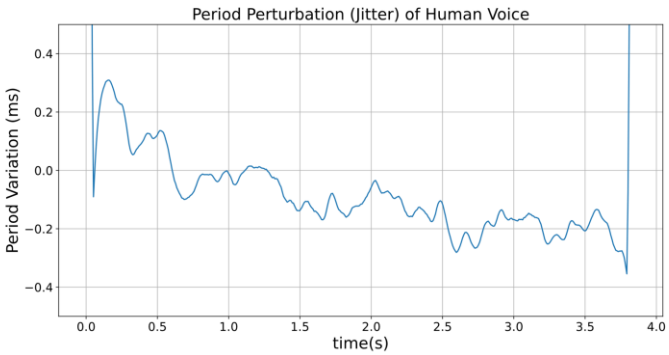
Sequencialmente foram sintetizadas 24 vogais /a/ sustentadas empregando-se modelo matemático proposto por Flanagan (FLANAGAN, 1965), em que sequência de pulsos no domínio da frequência, excitação E(z), é modulada por filtro glótico G(z) (ROSENBERG, 1969), sequencialmente por filtro de trato vocal V(z) (WAKITA, 1979) e por radiação de lábios L(z). As vogais foram sintetizadas com três diferentes condições de perturbação:

1. Perturbação humana – extraída de sinais de emissões gravadas por meio de correlação parcial.
2. Perturbação modelada – (a) segunda-ordem subamortecida acrescida de componente aleatório com distribuição uniforme e (b) apenas o módulo aleatório com distribuição uniforme.
3. Sem perturbação.

As vogais sintetizads foram randomizadas e avaliadas por três fonoaudiólogas experientes quanto ao grau de naturalidade - proximidade com a sonoridade da voz humana - utilizando escala analógica-visual de 100 mm, onde 0 significou ausência de naturalidade e 100 a máxima naturalidade. Comparou-se os graus de naturalidade resultantes com o teste estatístico Kruskal-Wallis e pós teste Dwass-Steel-Critchlow-Fligner.

RESULTADOS. A Figura 1 apresenta a representação gráfica da perturbação de período da voz humana. Observações desses traçados revelaram mais perturbação no início da emissão, assemelhando-se à resposta de uma segunda-ordem subamortecido. Além disso, oscilações foram observadas ao longo de toda a emissão, assemelhando-se a contornos gerados por efeitos aleatórios com uma distribuição aparentemente uniforme.

Figura 1: Perturbação de período presente na voz humana



As vogais sintetizadas encontram-se disponibilizadas na referência Davatz (2024). Os valores de média e desvio padrão do grau de naturalidade encontram-se apresentados na Tabela 1. Pode-se observar que vogais sem perturbação apresentaram escores baixos (2,2 ± 4,2). A inserção de perturbação usando os diferentes métodos aumentou significativamente a naturalidade, com diferença estatisticamente significativa em comparação à condição sem perturbação ($p < 0,001$). Além disso, o modelo de segunda-ordem com aleatoriedade (74,9 ±19,2) e o modelo aleatório com distribuição uniforme (68,6 ±24,9) demonstraram desempenho semelhante ($p=0,932$); e apesar de maior variabilidade dos valores, frequentemente se aproximaram da perturbação humana (86,5 ± 11,2) ($p= 0,271$ e $p=0,111$).

Tabela 1: Valores de media e desvio padrão do grau de naturalidade das vogais sintetizadas, em função do padrão de perturbação aplicado na síntese.

Categoria		Grau de Naturalidade							
		Perturbação Humana		2ª. Ordem		Aleatória		Sem perturbação	
Mulher	Jovem	84,9	±14,4	73,5	±20,8	54,4	±31,2	0,2	±0,3
	Meia-idade	87,0	± 6,0	73,7	±8,4	56,7	±7,6	0,1	±0,1
	Idosa	88,3	±5,7	61,3	±27,7	77,2	±20,4	0,0	±0,0
Homem	Jovem	87,2	±13,4	87,3	±9,1	84,1	±5,9	4,9	±3,8
	Meia-idade	81,5	±14,6	77,1	±11,7	53,1	±31,6	3,9	±5,3
	Idoso	90,0	±6,6	76,4	±19,5	86,0	±5,7	4,2	±6,0
	Total	86,5	±11,2	74,9	±19,2	68,6	±24,9	2,2	±4,2

*Segunda-ordem subamortecido acrescido de aleatoriedade.

CONSIDERAÇÕES FINAIS. A inspeção visual de traços gráficos de perturbação do período da voz humana permitiu a seleção de modelos matemáticos para a síntese de voz. O experimento demonstrou que introduzir perturbação no período dos sinais sintetizados usando modelos matemáticos que simulam as variações temporais no comprimento de onda da voz humana aumenta a naturalidade percebida.

PALAVRAS-CHAVE. Análise de voz. Jitter. Síntese de fala. Conceito de fonte-filtro.

AGRADECIMENTOS. Os autores agradecem às agências de fomento: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), sob o número 88887965017/2024-00, e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

AOKI, Naofumi; IFUKUBE, Tohru. Two 1/F fluctuations in sustained phonation and their roles on naturalness of synthetic voice. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON ELECTRONICS, CIRCUITS, AND SYSTEMS 1996, Anais [...]. [s.l: s.n.] p. 311–314. DOI: https://doi.org/10.1109/icecs.1996.582813.

CATALDO, Edson; SOIZE, Christian. Voice signals produced with jitter through a stochastic one-mass mechanical model. Journal of Voice, [S. l.], v. 31, p. 111.e9-111.e18, 2017. DOI: https://doi.org/10.1016/j.jvoice.2016.01.001.

DAVATZ, GC. Vogais sintetizadas sustentadas / a / com diferentes padrões de perturbação da frequência fundamental. 2024. Disponível em: https://drive.google.com/drive/folders/1JX_1QvCA69mgUlYEHYoD91iEJQuy3F0a?usp=sharing.

FARRÚS, M.; HERNANDO, J. Using jitter and shimmer in speaker verification. IET Signal Processing, [S. l.], v. 3, n. 4, p. 247–257, 2009. DOI: https://doi.org/10.1049/iet-spr.2008.0147.

FLANAGAN, James L. Speech analysis, synthesis and perception. 1aed. ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1965. DOI: https://doi.org/10.1007/978-3-662-00849-2.

KERSTA, L. G.; BRICKER, P. D.; DAVID, E. E. Human or machine?—a study of voice naturalness. The Journal of the Acoustical Society of America, [S. l.], v. 32, n. 11, p. 1502–1502, 1960. DOI: https://doi.org/10.1121/1.1935196.

ROSENBERG, A. E. Effect of glottal pulse shape on the quality of natural vowels. The Journal of the Acoustical Society of America, [S. l.], v. 46, n. 1A, p. 82–82, 1969. DOI: https://doi.org/10.1121/1.1912389. Disponível em: https://doi.org/10.1121/1.1912389.

TEIXEIRA, João Paulo; FERNANDES, Paula Odete. Jitter, Shimmer and HNR Classification within Gender, Tones and Vowels in Healthy Voices. Procedia Technology, [S. l.], v. 16, p. 1228–1237, 2014. DOI: 10.1016/J.PROTCY.2014.10.138.

WAKITA, Hisashi. Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art. IEEE Transactions on Acoustics, Speech, and Signal Processing, [S. l.], v. 27, n. 3, p. 281–285, 1979. DOI: https://doi.org/10.1109/TASSP.1979.1163242.