XXVIII Congresso {virtual} de Iniciação Científica da Unicamp 2020

Em busca dos arquétipos vocais: análise do jitter e das variações de frequência fundamental em diferentes categorias de falantes.

Aluno: Caio Henrique do Amaral – caioha25@gmail.com
Orientador: Prof. Dr. Tiago Fernandes Tavares
Coorientadora: Profa. Dra. Paula D. Paro Costa

I. Introdução

Traços afetivos, como o sentimento, a emoção e a personalidade de uma pessoa, podem estar ligados à sua maneira de falar. Uma voz trêmula, por exemplo, pode ser associada à insegurança de uma pessoa, ao passo que uma voz direta é, em geral, relacionada à assertividade. No aparelho fonador, a variação tonal da voz é controlada pela vibração da glote, e o número de ciclos glotais por segundo define a frequência fundamental (F0), ou pitch, da voz [1]. A variação da frequência fundamental em um sinal harmônico é chamada de *jitter* [1].

Podemos entender a frequência fundamental e o jitter como *features* acústicas, ou seja, características mensuráveis do sinal de áudio. A maneira como percebemos valores e variações de uma determinada feature está relacionada com sua a interpretação psicoacústica. Em nossa pesquisa, decidimos utilizar somente features que se relacionam a uma característica perceptual clara, visando uma explicação psicoacústica do resultado.

Neste projeto, verificamos a relevância do jitter e, posteriormente, de outras features acústicas na classificação de personalidades. Para isso, analisamos suas distribuições estatísticas em bases de dados. Observamos que o jitter carece de uma interpretação clara e não apresenta poder discriminativo, enquanto a frequência fundamental demonstra, quando comparada às features investigadas, o maior potencial classificatório.

II. Método

Inicialmente, extraímos os valores de jitter, calculados de acordo com [2], de uma base de dados com falas gravadas por atores [3]. Nesta base, os atores intencionalmente emulam as personalidades introvertido, balanceado e extrovertido. Em seguida, executamos o mesmo procedimento em uma base de dados com falas extraídas de vídeos na Internet, contendo alunos apresentando seus trabalhos de conclusão de curso (TCC), jornalistas e youtubers. Os alunos apresentando TCC deveriam relacionar-se à personalidade introvertida, os jornalistas, à neutra, e os youtubers, à extrovertida. A Fig. 1 mostra o número de arquivos de áudio em cada categoria e em cada base de dados.

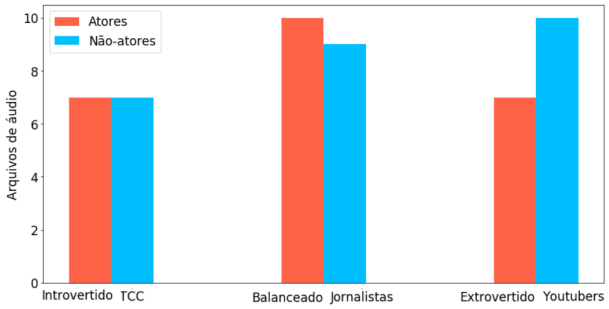

Fig. 1: Composição das bases de dados de atores e não-atores.

Os resultados da extração do jitter nas bases de dados, para as diferentes categorias, está mostrado na Fig. 2. Observamos que as categorias que considerávamos relacionadas não apresentavam distribuição estatística de jitter similares. De frente com esse obstáculo, realizamos experimentos em um estúdio profissional e descobrimos que o jitter não possui uma interpretação psicoacústica imediata. Além disso, contrastamos nosso resultados com estudos na área da linguística, que revelam a inaptidão do jitter em falas corridas [4] e em diferentes tipos de discurso [5]. Diante disso, passamos a analisar novas features na classificação de personalidades.

Em seguida, criamos uma nova base de dados de personagens da Disney, formada pelas arquétipos de heróis, princesas, vilões, vilãs e alívios cômicos. Os arquivos de áudio foram extraídos dos personagens dos filmes A Branca de Neve, Aladdin, Hércules, Mulan, Pequena Sereia, Rei Leão, Tarzan, A Bela e a Fera, Cinderela e A Bela Adormecida.

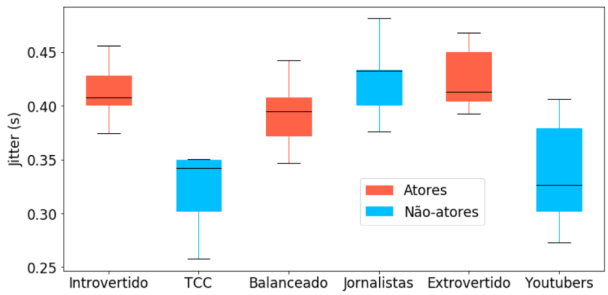

Fig. 2: Distribuição dos valores de jitter (s) das bases de dados de atores e não-atores.

A Fig. 3 detalha a quantidade de arquivos de áudio para cada arquétipo. Escolhemos personagens arquetípicos por transmitirem as suas características de forma mais direta, permitindo uma interpretação mais clara dos resultados. Na sequência, extraímos e analisamos a distribuição estatística da frequência fundamental, calculada com o auxílio da biblioteca Parselmouth [6], e de diversas features presentes na biblioteca Librosa [7].

III. Resultados

Ao analisarmos todas as features, percebemos que a frequência fundamental foi a que apresentou maior relação custo-benefício entre a capacidade de classificar os arquétipos e interpretação psicoacústica dos resultados. Isto é, a F0 ostenta maior distribuição estatística entre os diferentes arquétipos da base de dados e possui relação interpretável entre as sensações auditivas e as características físicas da fala. Sua distribuição é apresentada na Fig. 4.

IV. Conclusão

Podemos conjecturar uma relação entre a distribuição estatística dos valores de F0 com a caracterização dos arquétipos nos filmes. Os heróis apresentam um pitch grave com poucas variações, indicando uma fala mais assertiva e segura. Já os vilões e vilãs expressam uma variação um pouco maior, o que podemos entender como uma dissimulação na fala. As princesas e os alívios cômicos manifestam intensas variações e pitch mais altos quando comparados as demais categorias, características que, quando atreladas, associamos à insegurança e à quebra de expectativa.

O poder discriminativo da frequência fundamental abre espaço para a criação de sistemas de feedback automático e dinâmico para pessoas em processo de treinamento de locução. Além disso, aponta que a frequência fundamental é uma característica acústica de grande importância em estudos que utilizam a fala na área da computação afetiva.

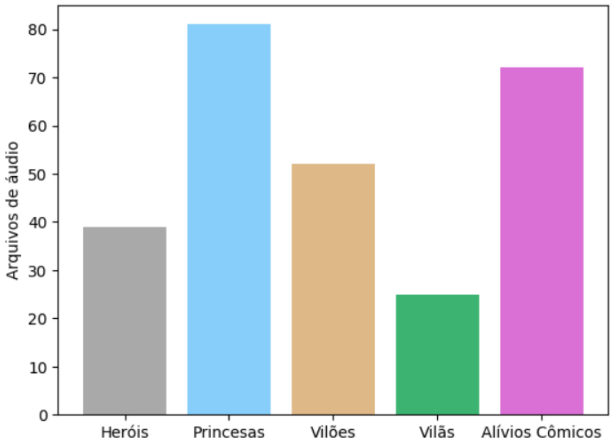

Fig. 3: Composição da base de dados de arquétipos da Disney.

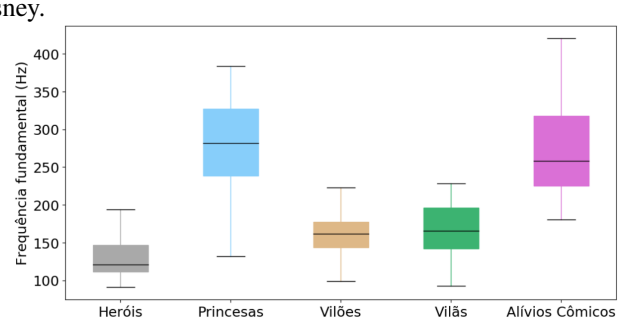

Fig. 4: Distribuição dos valores de frequência fundamental (Hz) da base de dados de arquétipos.

References

[1] O. P. P. A. d. L. Behlau, Mara Suzana; Tosi, "Determinação da frequência fundamental e suas variações em altura 'jitter' e intensidade 'shimmer', para falantes do português brasileiro," 1985.

[2] B. W. S. J. S. E. A. C. B. L. Y. D. J. E. P. L. S. S. N. K. P. T. Florian Eyben, Klaus R. Scherer, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, 2016.

[3] P. D. P. Costa, *Two-dimensional expressive speech animation*. PhD thesis, PhD thesis, School of Electrical and Computer Engineering, 2015.

[4] P. A. Jean Schoentgen, "Analysis and synthesis of vocal flutter and vocal jitter," *INTERSPEECH 2019*, 2019.

[5] P. A. SIQUEIRA, Júlia e BARBOSA, "Diferencas prosodicas em atos combinados a atitudes distintas," *Anais do I Congresso Brasileiro de Prosódia*.

[6] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[7] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.