Wpływ indukowanego bodźca afektywnego na właściwości fizyczne mowy ludzkiej

Opiekun projektu badawczego

dr Krzysztof Basiński

Autorzy

Tomasz Domżalski

Abstrakt

Wstęp

Ekspresja emocji jest istotnym elementem komunikacji interpersonalnej oraz nośnikiem informacji o procesach wewnątrzpsychicznych. Mogą być wyrażane z wykorzystaniem komunikacji werbalnej, zachowania, czy też komunikacji niewerbalnej. Wiele mierzalnych procesów fizjologicznych takich jak tętno, reakcja elektrodermalna, potencjał czynnościowy inaczej impuls nerwowy, mogą być wyznacznikiem eksternalizacji procesów emocjonalnych (Swain & Routray, 2018). Jednak, najpowszechniej dostępnym medium stanów afektywnych jest głos. Zagadnienie identyfikacji emocji w oparciu o sygnał mowy jest problemem interdyscyplinarnym, który angażuje nauki medyczne, społeczne oraz techniczne.

Wraz z rozwojem technologii, opracowywane są automatyczne systemy rozpoznające emocje w głosie mówcy (Igras & Wszolek, 2012).Takie rozwiązania mają możliwie szerokie zastosowania – (1) system bezpieczeństwa kodujący informacje o stanie psychicznym kierowcy do systemu operacyjnego w samochodzie, (2) narzędzia diagnostyczne, (3) tłumaczenie języka w czasie rzeczywistym z uwzględnieniem zabarwienia emocjonalnego wypowiedzi (El, Kamel, & Karray, 2011). Ocena nasilenia zaburzeń depresyjnych oraz skuteczności terapi (Mundt, Snyder, Cannizzaro, Chappie, & Geralts, 2007). Predykcja wystąpienia zaburzeń psychotycznych na podstawie okresu występowania objawów prodromalnych (Rezaii, Walker, & Wolff, 2019). Automatyczna klasyfikacja osób z zaburzeniami depresyjnymi i osób z ryzykiem popełnienia samobójstwa w oparciu o analizę cepstralną dźwięków wytworzonych przez głośnię (Risk et al., 2004).

Uzyskanie reprezentatywnej próbki mowy w warunkach naturalnych jest trudne. Różne bazy danych składają się z nagrań z telewizji, programów radiowych, rozmów z numerem alarmowym (Kaminska, Sapinski, & Pelikant, 2013). Funkcjonują również sztucznie przygotowane zestawy, gdzie aktorzy lub wolontariusze, czytają nacechowane emocjonalnie zdania. Kolejnym wariantem

są nagrania z indukowanymi emocjami w warunkach eksperymentalnych (Igras & Ziólko, 2009).

Dotychczasowe badania w obszarze identyfikacji cech fali akustycznej skorelowanych ze stanem emocjonalnym, wskazuje się na parametry związane z energią sygnału, współczynniki MFCC (Mel-frequency cepstral coefficients) oraz częstotliwość podstawowa (f0). Ta ostatnia, inaczej nazywana tonem krtaniowym, charakteryzuje się stosunkowo wysoką niezależnością od szumów napływających z otoczenia (Dimitrova-Grekow, Klis, & Igras-Cybulska (2019); Igras & Wszolek (2012)). Właściwością tego parametru jest rozróżnienie częstotliwości podstawowej względem płci, gdzie kobiety posługują się nią w wyższym zakresie niż mężczyźni (Selvaraj, Bhuvana, & Padmaja, 2016). Istotnym czynnikiem, który należy wziąć pod uwagę są zmiany w wartościach dla niektórych parametrów w zależności od wieku badanego. Częstotliwość podstawowa wraz z wiekiem rośnie u mężczyzn, a maleje u kobiet (Ferrand, 2002). Pozytywnie nacechowana wypowiedź charakteryzuje się większą rozpiętoością częstotliwości podstawowej oraz jest głośniejsza. Natomiast ta z negatywnym ładunkiem emocjonalnym cechuje się mniejszą rozpiętością wartości, które przyjmuje f0 i jest cichsza (Globerson, Amir, Golan, Kishon-Rabin, & Lavidor, 2013). Istotnym parametrem dla ekspresji emocji w głosie jest również parametr HNR(Harmonics-to-Noise Ratio). Niska wartość tej cechy sprawia, że głos odbierany jest percepcyjnnie jako ochrypły. Im wyższa jest jego wartość, tym głos odbierany jest jako czystszy (Hakanpää, Waaramaa, & Laukkanen, 2019) .

Celem tego badania jest analiza wybranych parametrów akustycznych mowy dla eksperymentalnie indukowanych stanów afektywnych u uczestników badania. Analiza ma posłużyć weryfikacji czy wybrane cechy sygnału mowy będą wystarczające do rozróżnienia pozytywnego afektu od negatywnego. Hipoteza badawcza zakłada wystąpienie różnic wywołanych bodźcem emocjonalnym.

Metody

Badani:

W celu uniknięcia różnic wynikających z płci, w badaniu z powtarzanym pomiarem wzięły udział 23 kobiety w wieku M = 24.04 (SD = 1.82). Warunkami dopuszczającym do udziału w badaniu był status studentki oraz nie wykonywanie pracy głosem (np. aktor, lektor, speaker radiowy).

Bodziec afektywny:

Do wywołania reakcji emocjonalnej wykorzystano wystandaryzowaną bazę zdjęć na licencji open – access(Kurdi et al., 2017). Wyselekcjonowanych zostało 30 zdjęć, po 10 zdjęć dla każdego z warunków. Wybrano zdjęcia z najniższym możliwym odchyleniem standardowym dla walencji i uśrednionym pobudzeniem.

Dla warunku neutralnego(neut) - walencja M = 4.02 (SD = 0.05) oraz pobudzeniem M = 1.79 (SD = 0.11). Dla warunku negatywnego(neg) – walencja M = 1.63 (SD = 0.17) oraz pobudzenie M = 4.5 (SD = 0.27). Dla warunku pozytywnego(pos) – walencja M = 6.51 (SD = 0.07) oraz pobudzenie M = 4.55 (SD = 0.35).

Materiał akustyczny:

Materiał akustyczny zawierał przeczytane przez badanego 10 afektywnie neutralnych zdań wybranych na podstawie wcześniejszej opublikowanej pracy (Ben-David, Van Lieshout, & Leszcz, 2011). Zdania w oryginale były w języku angielskim. Na potrzebę badania zostały one przetłumaczone na język polski, a ich neutralność afektywna sprawdzona za pomocą badania pilotażowego. Badanie polegało na ocenie przez uczestników natężenia emocjonalnego zdań w skali Likerta, od 1 – zdecydowanie nie, do 5 – zdecydowanie tak. W badaniu pilotażowym wzięło udział 28 osób w wieku M = 23.57 (SD = 2.04). Wynik średni dla zdań wyniósł M = 1.40 (SD = 0.20).

Parametry sygnału:

Ekstrakcja cech została przeprowadzona z wykorzystaniem biblioteki Surfboard stworzonej w języku programowania Python (Lenain, Weston, Shivkumar, & Fristed, 2020) . Częstotliwość próbkowania dla przebiegu załadowanych fal akustycznych wynosiła 48kHz. Wyselekcjonowane cechy to: (1) średnia częstotliwość podstawowa (f0_mean) – związek z intonacją i stanem emocjonalnym człowieka (Wilk, 2015) , (2) średnia wartość RMS (Root-Mean-Square) – parametr odpowiadający energii sygnału (Ahsan & Kumari, 2016) (3) ang. Spectral Centroid – parametr mówiący o barwie głosu, ale też kształcie spectrum dźwięku (Chatterjee, Mukesh, Hsu, Vyas, & Liu, 2018), (4) HNR – pozwala określić wydajność mowy, inaczej energię wibracji strun głosowy wywołanych przepływem powietrza z płuc (Teixeira, Oliveira, & Lopes, 2013).

Procedura:

Procedura została zbudowana i przeprowadzona z wykorzystaniem graficznego interfejsu lab.js opartego na języku programowania JavaScript (Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, 2020). Uczestnicy badania po otwarciu linku zapoznawali się z warunkami przystąpienia do badania, przejść dalej mogli jedynie po ich zaakceptowaniu. Po przeczytaniu instrukcji uruchamiała się sekwencja zawierająca 10 zdjęć, każde wyświetlane było przez 3 sekundy I zmieniały się automatycznie. Po zakończeniu wyświetlania zdjęć, na ekranie pojawiało się 10 zdań, które badany miał przeczytać na głos. W tym czasie uczestnik był rejestrowany poprzez mikrofon dostępny

w urządzeniu badanego, zgodnie z instrukcją - w słuchawkach. Po przeczytaniu zdań, badany poprzez naciśnięcie przycisku spacja przechodził do sekwencji ze zdjęciami dla kolejnego warunku. Kolejność zdjęć I warunków była losowa dla każdego uczestnika. Po obejrzeniu 3 sekwencji zdjęć, gdzie po każdej badany miał do przeczytania te samo 10 zdań w niezmienionej kolejności, badany pobierał nagrania i wysyłał je na podany adres e-mail badacza.

Statystyka:

Analiza danych została przeprowadzona z wykorzystaniem bibliotek dla języka programowania Python: NumPy oraz Pandas (team, 2020; Van Der Walt, Stefan and Colbert, S Chris and Varoquaux, 2011), analiza statystyczna, w tym ANOVA jednoczynnikowa dla powtarzanych pomiarów, została wykonana za pomocą biblioteki Statsmodels(Seabold & Perktold, 2010), wizualizacje powstały z wykorzystaniem bilbioteki Seaborn (Waskom, Botvinnik, & O'Kane, 2017).

Wyniki

Analiza wariancji wskazała na brak istotnego efektu głównego dla wszystkich branych pod uwagę cech. W przypadku średniej częstotliwości podstawowej, $F(2, 44) = 0.74$, $p = .48$. Dla hnr, $F(2, 44) = 0.75$, $p = .48$. Dla parametru rms_mean, $F(2, 44) = 1.88$, $p = .16$, a dla cechy spectral_centroid_mean, $F(2, 44) = 1.99$, $p = .14$.

Średnie wartości dla parametrów f0, spectral_centroid_meancentroid oraz hnr są najwyższe dla warunku stymulacji pozytywnej. Dla cech f0_mean, hnr oraz rms_mean, wyniki średnie są njaniższe dla warunku neutralnego. Średnia wartość parametru rms jest najwyższa dla warunku stymulacji negatywnej (Tabela 1).

Tabela 1. Średnie wartość oraz odchylenie standardowe parametrów dla każdego z warunków.

condition	f0_mean	SD	hnr	SD	rms_mean	SD	spectral_centroid_mean	SD
neg	197.445	13.16	11.6749	1.48	0.061282	0.03	3484.65	1099.31
neut	197.257	13.20	11.641	1.66	0.0588899	0.04	3571.36	1167.7
pos	198.388	11.50	11.8001	1.81	0.0590017	0.03	3604.21	1203.43

Mediana wartości dla parametru f0_mean jest wyższa dla warunku stymulacji pozytywnej(Mdn = 200.647) od neutralnego (Mdn = 198.414) oraz negatywnego (Mdn = 197.872) (Fig. 1). Dla parametru hnr najwyższa wartość występuje dla warunku pozytywnego (Mdn = 11.794), warunek neutralny (Mdn = 11.682), warunek negatywny (Mdn = 11.5976) (Fig. 3). Cecha rms_mean ma medianę warunku pozytywnego (Mdn = 0.069) wyższą od warunku neutralnego (Mdn

= 0.066) oraz negatywnego (Mdn = 0.064) (Fig. 2). Mediana parametru spectral_centroid_mean dla warunku negatywnego (Mdn = 3571.39) jest wyższa od pozytywnego (Mdn = 3420.13) i neutralnego (Mdn = 3227.42) (Fig. 4).

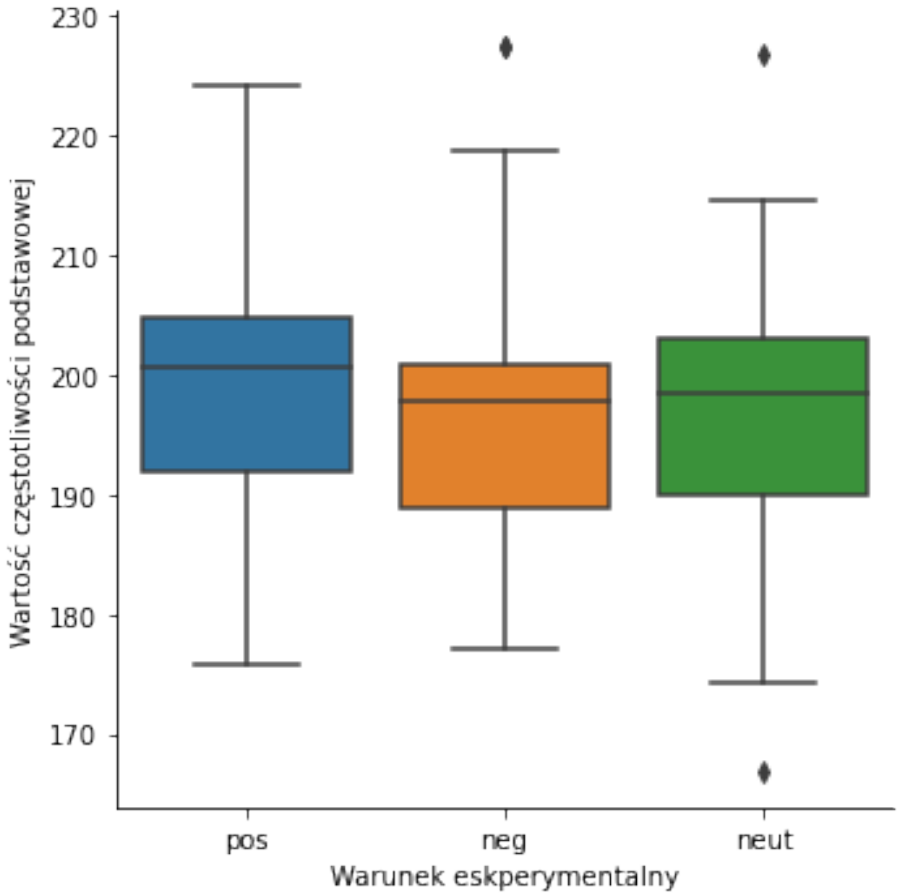

Figure 1: Fig.1: Wartości f0_mean względem warunków eksperymentalnych

Dyskusja

Analiza statystyczna wykazała brak istotnych różnic między grupami. Wynik ten jodnoznacznie nie potwierdza postawionej hipotezy. Miara średniej jest wrażliwa na skośnośc rozkładu danych jak i występowanie skrajnych wyników. Obydwa czynniki wystąpiły w przypadku otrzymanych wyników. Mediana w tym wypadku jest lepszą miarą porównawczą. W przypadku częstotliwości podstawowej można dostrzec nieco wyższą mediane dla warunku pozytywnego, natomiast dla negatywnego niższą. Sugeruje to, że odczuwając pozytywne

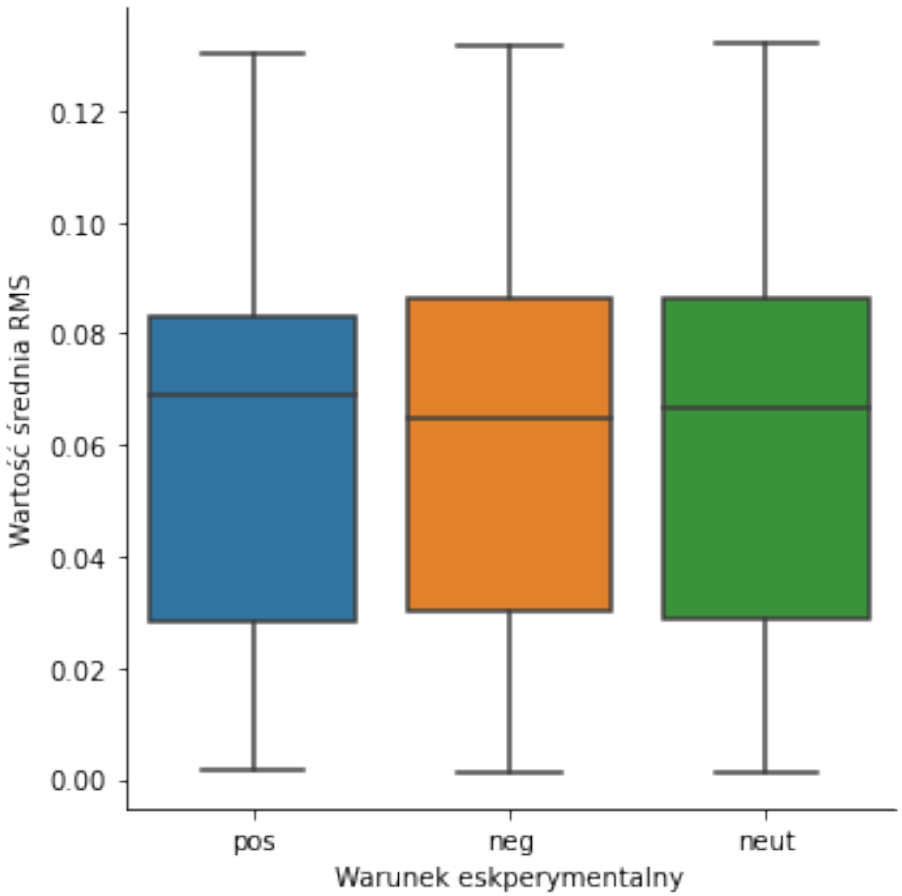

Figure 2: Fig.2: Wartości rms_mean względem warunków eksperymentalnych

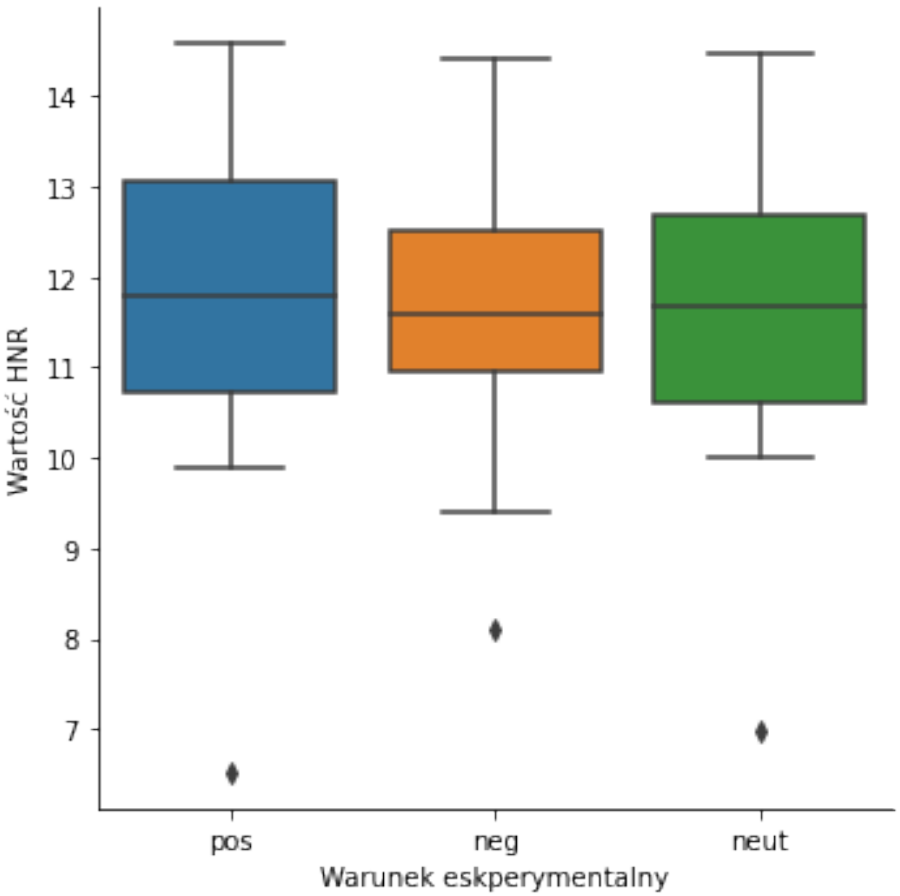

Figure 3: Fig.3: Wartości hnr względem warunków eksperymentalnych

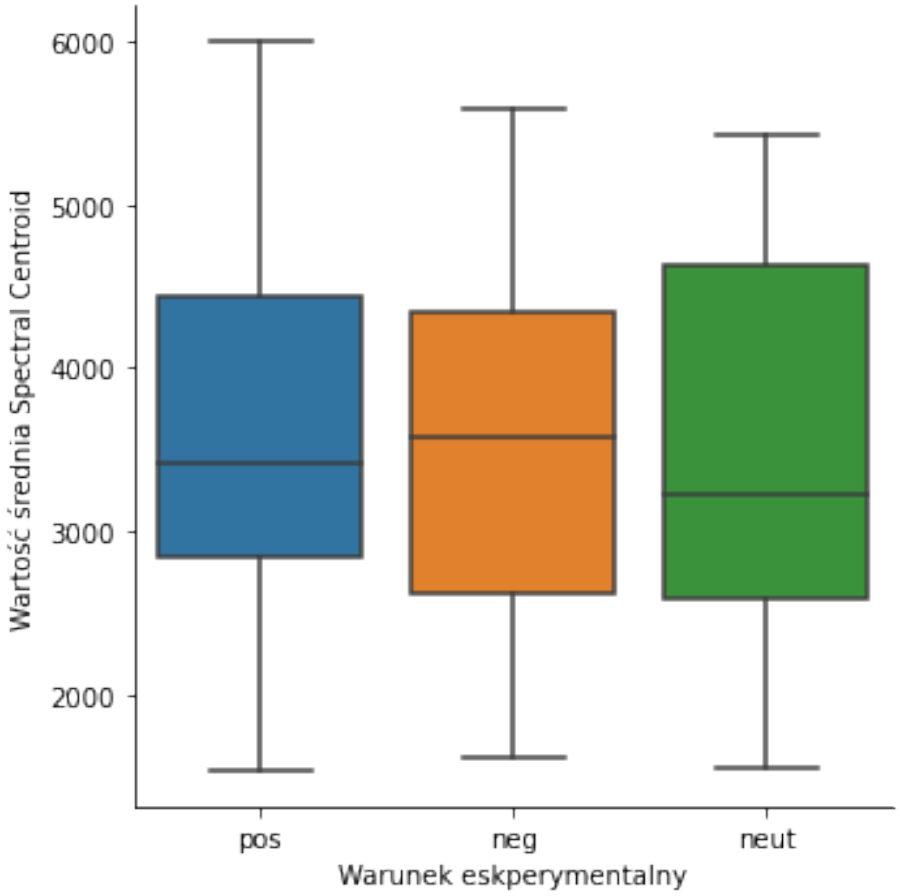

Figure 4: Fig.4: Wartości spectral_centroid_mean względem warunków eksperymentalnych

emocje mówimy nieco wyżej, a gdy odczuwamy nieprzyjemne emocje, mówimy niżej. Mediana dla parametru spectral_centroid_mean można interpretować jako częstsze występowanie pauz w warunku negatywnym. Mimo braku istotnych różnic można dojrzeć tendencję kierunku, w którym te różnice zmierzają. Dla cech f0, rms oraz hnr są to wyższe wyniki dla stymulacji pozytywnej oraz najniższe dla negatywnej.

Czynnikami, które mogły wpłynąć na wyniki w sposób niekorzystny to manipulacja eksperymentalnea, która mogła być silniejsza. Również, możliwym jest, że uczestnicy automatyzowali czynność czytania zdań, których było 10 i wyświetlane były zawsze w tej samej kolejności - redukowało to wpływ stymulacji afektywnej na sygnał mowy. Materiał akustyczny był nagrywany w niekontrolowanych eksperymentalnie warunkach - badani realizowali procedurę z wykorzystaniem własnego urządzenia, w warunkach prawdopodobnie domowych. Aby zniwelować wpływ tego czynnika należałoby nagrywać uczestników dyktafonem w tych samych warunkach - najlepiej w wyizolowanym akustycznie pomieszczeniu. Ekstrakcja większej ilości parametrów mogłaby dać lepsze rezultaty. Parametry oparte na analizie spektrum dźwięku takie jak MFCC czy LFPC (Log Frequency Power Coefficients) są często wykorzystywane dla klasyfikacji maszynowej 6 podstawowych emocji, dając pozytywne rezultaty (Liang, Zheng, & Zeng, 2019).

Wyniki tego badania nie są wystarczająco konkluzywne, aby uznać wybrane parametry jako nieodpowiednie do rozrózniania stanów afektywnych w głosie człowieka. Uwzględniając poprawki w manipulacji eksperymentalnej, metodzie poboru materiału akystycznego oraz procedurze czytania zdań - badanie może dać rozsztrygające rezultaty.

Bibliografia:

Ahsan, M., & Kumari, M. (2016). Physical Features Based Speech Emotion Recognition Using Predictive Classification. *International Journal of Computer Science and Information Technology*, *8*(2), 63–74. https://doi.org/10.5121/ijcsit.2016.8205

Ben-David, B. M., Van Lieshout, P. H., & Leszcz, T. (2011). A resource of validated affective and neutral sentences to assess identification of emotion in spoken language after a brain injury. *Brain Injury*, *25*(2), 206–220. https://doi.org/10.3109/02699052.2010.536197

Chatterjee, J., Mukesh, V., Hsu, H. H., Vyas, G., & Liu, Z. (2018). Speech emotion recognition using cross-correlation and acoustic features. *Proceedings - IEEE 16th International Conference on Dependable, Autonomic and Secure Computing, IEEE 16th International Conference on Pervasive Intelligence and Computing, IEEE 4th International Conference on Big Data Intelligence and Computing and IEEE 3*, (October), 250–255. https://doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00050

Dimitrova-Grekow, T., Klis, A., & Igras-Cybulska, M. (2019). Speech emotion recognition based on voice fundamental frequency. *Archives of Acoustics*, *44*(2), 277–286. https://doi.org/10.24425/aoa.2019.128491

El, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition : Features , classification schemes , and databases. *Pattern Recognition*, *44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Ferrand, C. T. (2002). Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice*, *16*(4), 480–487. https://doi.org/10.1016/S0892-1997(02)00123-6

Globerson, E., Amir, N., Golan, O., Kishon-Rabin, L., & Lavidor, M. (2013). Psychoacoustic abilities as predictors of vocal emotion recognition. *Attention, Perception, and Psychophysics*, *75*(8), 1799–1810. https://doi.org/10.3758/s13414-013-0518-x

Hakanpää, T., Waaramaa, T., & Laukkanen, A. M. (2019). Comparing Contemporary Commercial and Classical Styles: Emotion Expression in Singing. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2019.10.002

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2020). *ab.js: A free, open, online study builder.* https://doi.org/10.5281/zenodo.597045

Igras, M., & Wszolek, W. (2012). Pomiary parametrów akustycznych mowy emocjonalnej - krok ku modelowaniu wokalnej ekspresji emocji. *Pomiary Automatyka Kontrola*, *R. 58, nr*(4), 335–338.

Igras, M., & Ziółko, B. (2009). Baza Danych Nagrań Mowy Emocjonalnej Database of Emotional Speech Recordings. *Studia Informatica*, *30*(182).

Kaminska, D., Sapinski, T., & Pelikant, A. (2013). Review of Emotion Recognition from Speech. *Researchgate.Net*, (January 2015). Retrieved from https://www.researchgate.net/profile/Dorota{_}Kaminska/publication/266968747{_}Review{_}of{_}E

Lenain, R., Weston, J., Shivkumar, A., & Fristed, E. (2020). *Surfboard: Audio Feature Extraction for Modern Machine Learning*. 1–5. Retrieved from http://arxiv.org/abs/2005.08848

Liang, Y., Zheng, X., & Zeng, D. D. (2019). A survey on big data-driven digital phenotyping of mental health. *Information Fusion*, *52*(March), 290–307. https://doi.org/10.1016/j.inffus.2019.04.001

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics*, *20*(1), 50–64. https://doi.org/10.1016/j.jneuroling.2006.04.001

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *Npj Schizophrenia*, *5*(1). https://doi.org/10.1038/s41537-019-0077-9

Risk, N.-t. S., Ozdas, A., Shiavi, R. G., Member, S., Silverman, S. E., Silverman, M. K., & Wilkes, D. M. (2004). *Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and*. *51*(9), 1530–1540.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proc. Of the 9th Python in Science Conf*, (Scipy), 92–96. Retrieved from http://statsmodels.sourceforge.net/

Selvaraj, M., Bhuvana, R., & Padmaja, S. (2016). Human speech emotion recognition. *International Journal of Engineering and Technology*, *8*(1), 311–323. https://doi.org/10.1145/3129340

Swain, M., & Routray, A. (2018). Databases , features and classifiers for speech emotion recognition : a review. *International Journal of Speech Technology*, *0*(0), 0. https://doi.org/10.1007/s10772-018-9491-z

team, T. pandas development. (2020). *pandas-dev/pandas: Pandas*. https://doi.org/10.5281/zenodo.3509134

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology*, *9*, 1112–1122. https://doi.org/10.1016/j.protcy.2013.12.124

Van Der Walt, Stefan and Colbert, S Chris and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science \& Engineering*, *13*(2), 22.

Waskom, M., Botvinnik, O., & O'Kane, D. (2017). *Mwaskom/seaborn: V0.8.1*. https://doi.org/10.5281/zenodo.883859

Wilk, B. (2015). Wyznaczanie wartości chwilowej cz stotliwości podstawowej tonu krtaniowego za pomoca analizy falkowej sygnalu mowy. *Przeglad Elektrotechniczny*, *91*(11), 305–308. https://doi.org/10.15199/48.2015.11.70