Rozpoznawanie mowy w obecności zakłóceń

Rafał Sokołowski
Opiekun: dr Marek Skarupski

Wydział Matematyki
Politechnika Wrocławska

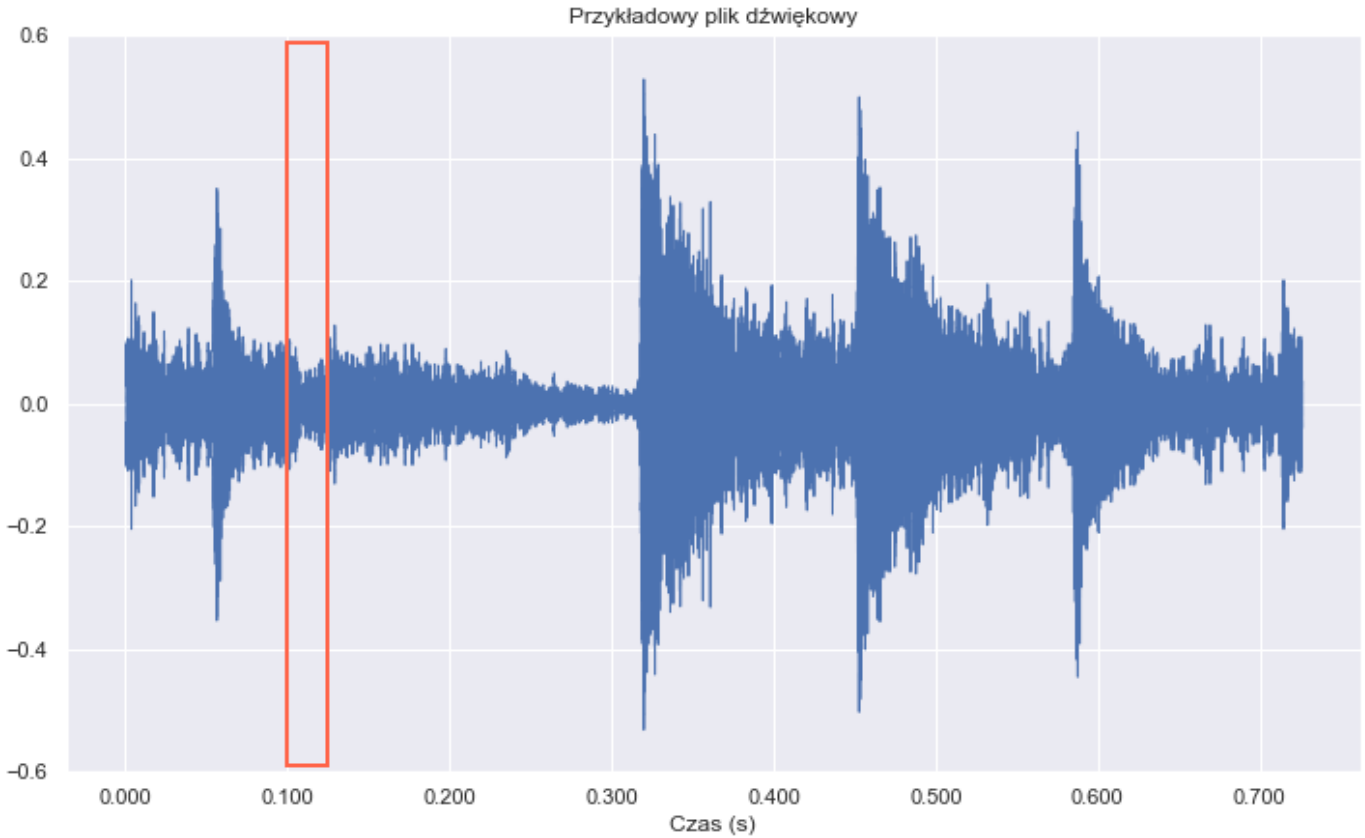
20 stycznia 2022

Spis treści

Czym jest rozpoznawanie mowy?

Rysunek: Źródło: https://developer.nvidia.com/

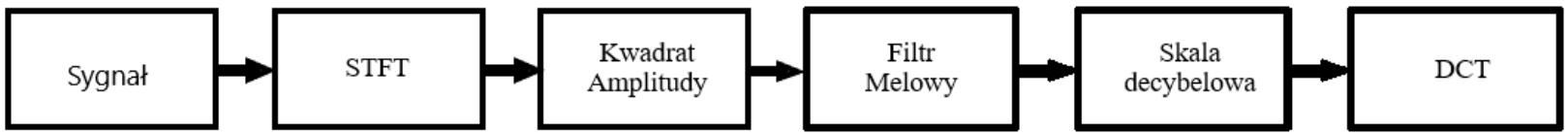
Dane akustyczne



Rysunek: Źródło: Opracowanie własne

Przygotowane danych - Współczynniki MFCC

Współczynniki MFCC (ang. Mel-frequency cepstrum coefficients, [3]) stanowią reprezentację widma M najistotniejszych zakresów częstotliwości sygnału dźwiękowego.



Rysunek: Źródło: Opracowanie własne

Ukryte Łańcuchy Markowa - własność Markowa

Definicja

Jeśli proces stochastyczny $\{q_t, t \in \mathcal{T}\}$ spełnia tzw. własność Markowa, czyli

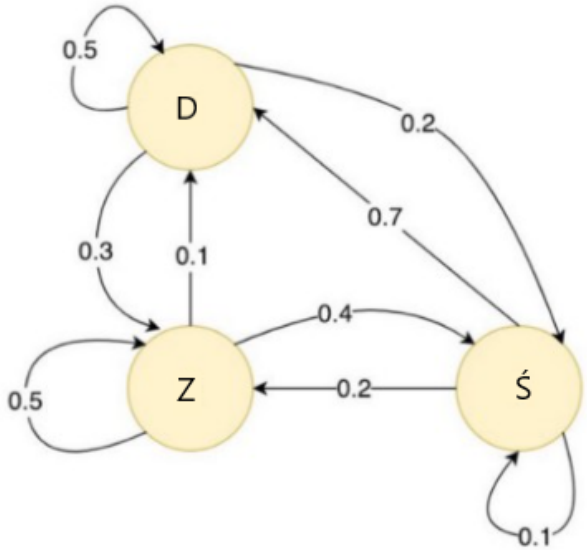
$$P(q_{t+1} = i | q_t, q_{t-1}, \ldots, q_0) = P(q_{t+1} = i | q_t) \tag{1}$$

dla i pochodzącego z przestrzeni stanów, to mówimy, że proces jest **łańcuchem Markowa pierwszego rzędu**.

Ukryte Łańcuchy Markowa - łańcuchy Markowa

Definiujemy prawdopodobieństwo przejścia

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i) \qquad \forall t \in \mathcal{T}, \forall s_i, s_j \in \{\mathrm{D}, \mathrm{Z}, \acute{\mathrm{S}}\}. \tag{2}$$



Łańcuch Markowa

	S_1	S_2	S_3
S_1	0.5	0.1	0.7
S_2	0.3	0.5	0.2
S_3	0.2	0.4	0.1

Macierz przejścia

Rysunek: Źródło: https://jonathan-hui.medium.com/

Ukryte Łańcuchy Markowa

W przypadku ukrytych łańcuchów Markowa (HMM) mamy do czynienia z dwoma procesami stochastycznymi:

- $\{q_t, t \in \mathcal{T}\}$ przyjmujący wartości z przestrzeni stanów ukrytych Q.
- $\{O_t, t \in \mathcal{T}\}$ przyjmujący wartości z przestrzeni stanów jawnych $\mathcal{O}$.

Ukryte Łańcuchy Markowa

Dla procesu q_t mamy spełnioną własność Markowa, czyli

$$P(q_{t+1} = i|q_t, q_{t-1}, \ldots, q_0) = P(q_{t+1} = i|q_t) \quad \forall i \in Q \tag{3}$$

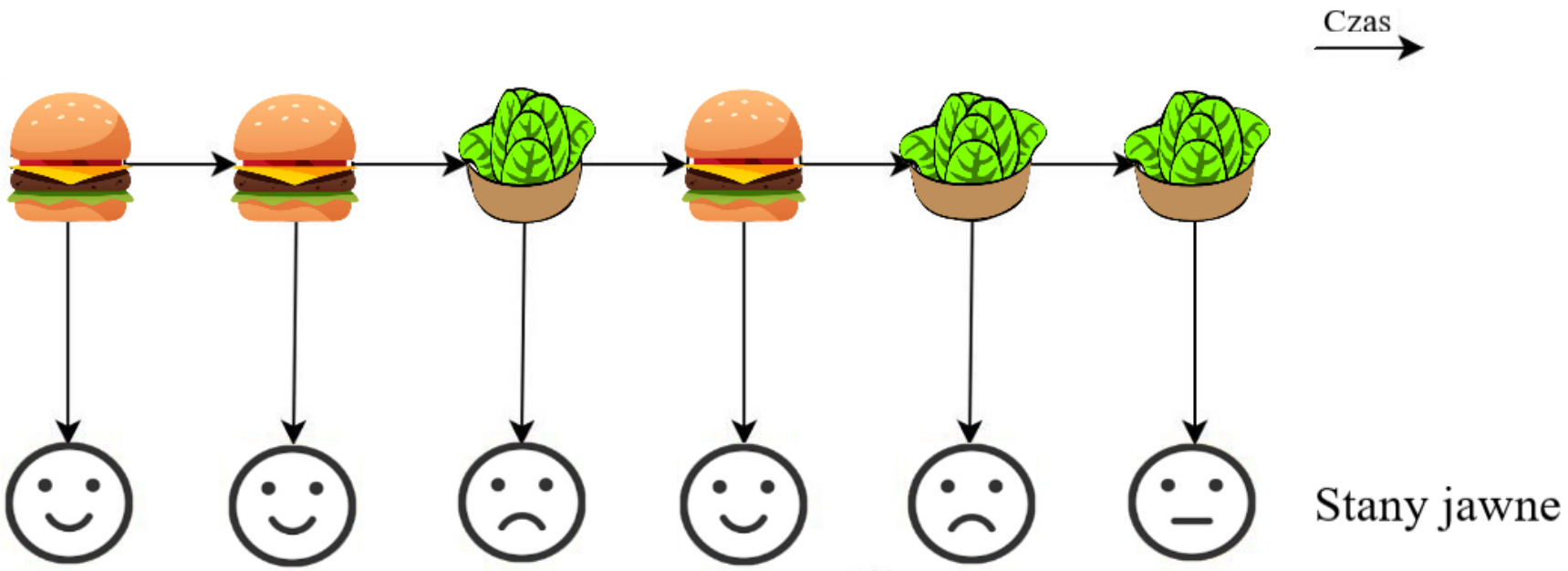
Natomiast dla O_t, mamy spełniony warunek niezależności wyniku z momentu t od poprzednich obserwacji, tzn $\forall i \in Q, \forall j \in \mathcal{O}$:

$$P(O_t = j|q_t, \ldots, q_0, O_{t-1}, \ldots, O_0) = P(O_t = j|q_t = i). \tag{4}$$

Ukryte Łańcuchy Markowa - przykład

Definiujemy prawdopodobieństwo emisji:

$$b_{ij} = P(O_t = j | q_t = i) \qquad \forall i \in Q, \forall j \in \mathcal{O}. \tag{5}$$



Rysunek: Źródło: Opracowanie własne

Ukryte Łańcuchy Markowa

Zdefiniujmy wektor stanu początkowego $\pi = \{\pi_i\}$, $i \in Q$, gdzie

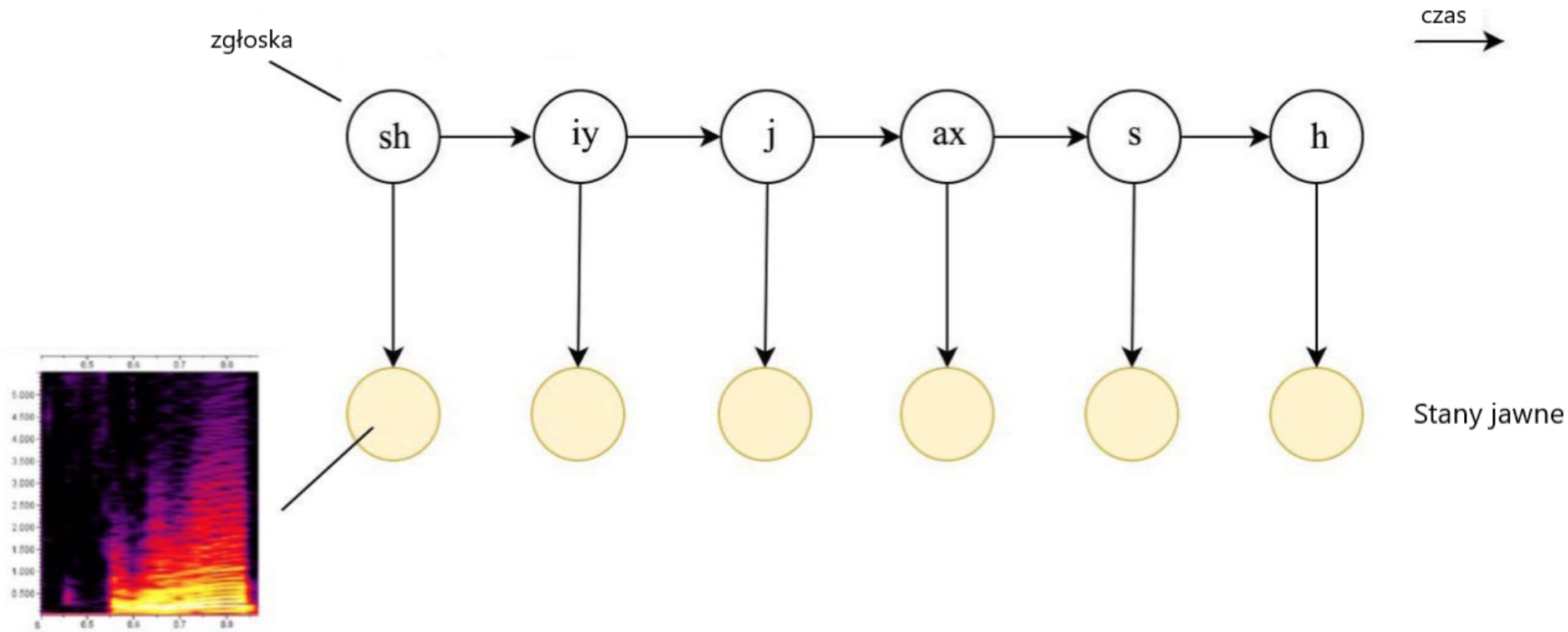
$$\pi_i = P(q_0 = i) \tag{6}$$

Wtedy krotka $\lambda = (Q, \mathcal{O}, \mathrm{A}, \mathrm{B}, \pi)$, gdzie

- $\mathrm{A} = \{a_{ij}\}$ - macierze prawdopodobieństw przejścia
- $\mathrm{B} = \{b_{ij}\}$ - macierze prawdopodobieństw emisji

Jednoznacznie wyznacza ukryty łańcuch Markowa (p. [5], [6]).

Ukryte Łańcuchy Markowa w rozpoznawaniu mowy



Rysunek: Źródło: https://jonathan-hui.medium.com

Szumy Gaussowskie

Rozpatrzmy sygnał $\{N_t : t = 0, \pm 1, \pm 2, \dots\}$ jako proces gaussowski o średniej 0, tzn.

$$E[N_t] = 0 \quad \forall t \in \mathbb{R}. \tag{7}$$

Zdefiniujemy funkcję autokowariancji procesu N_t jako

$$r(h) = Cov[N_t, N_{t-h}] = E[N_t \cdot N_{t-h}] \quad \forall t, h \in \mathbb{R}. \tag{8}$$

Definicja

Niech proces stochastyczny N_t spełnia (7) i (8). Widmowa gęstość mocy $\phi(f)$ (ang. *Power Spectrum Density*, PSD, [7]) zdefiniowana jest jako krótkoczasowa dyskretna transformata Fouriera funkcji autokowariancji $r(h)$, tzn.

$$\phi(f) = \sum_{h=-\infty}^{\infty} r(h) e^{-i \cdot f \cdot h}, \tag{9}$$

gdzie f to częstotliwość.

Szumy Gaussowskie

Będziemy rozważać gaussowskie szumy losowe, będące procesami stochastycznymi N_t, których widmowa gęstość mocy $\phi(f)$ będzie proporcjonalna do wartości $(1/f)^\beta$. W szczególności (p. [1], [8]):

Definicja

- $\beta = -2$, $\phi(f) \propto f^2$ – fioletowy szum.
- $\beta = -1$, $\phi(f) \propto f$ – niebieski szum.
- $\beta = 0$, $\phi(f)$ jest stała – biały szum.
- $\beta = 1$, $\phi(f) \propto f^{-1}$ – różowy szum.
- $\beta = 2$, $\phi(f) \propto f^{-2}$ – czerwony szum, znany również jako szum brownowski.

Zbiór danych

W celu uzyskania idealnego zbioru danych wykorzystamy narzędzie z biblioteki pythona *gTTS* (p. [2]). Umożliwi nam on wygenerowanie sygnałów dźwiękowych od zadanych klas słów:

id	Słowo
0	down
1	go
2	left
3	no
4	off
5	on
6	right
7	stop
8	to
9	yes

Tabela: Rozpatrywane klasy słów. Źródło: Opracowanie własne

Współczynnik SNR

Definicja

Moc sygnału $\{S_t, t \in \mathcal{T}\}$, który jest procesem stochastycznym definiujemy jako:

$$P = E|S_t|^2 \quad \forall t \in \mathcal{T}, \tag{10}$$

czyli jest to drugi moment rozkładu S_t.

Do zmierzenia poziomu zaszumienia sygnału wykorzystamy współczynnik SNR (ang. *Signal-noise ratio*, [4]) zdefiniowany następująco:

$$SNR = \frac{P_S}{P_N}, \tag{11}$$

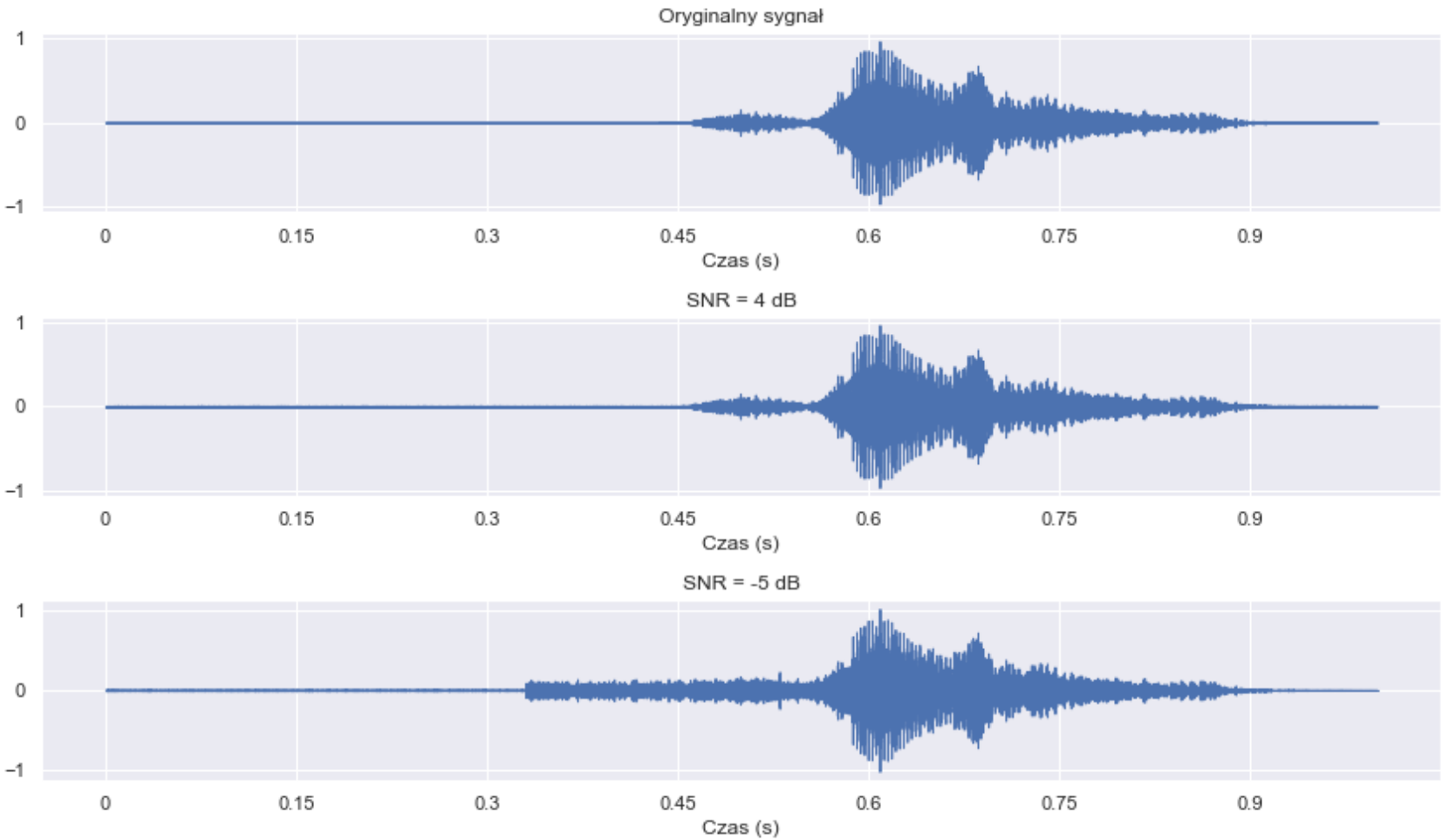
gdzie P_S, P_N to odpowiednio moc sygnału i moc szumu.

Współczynnik SNR

Wygodniej będzie nam przedstawić równanie (11) w skali decybelowej, uzyskujemy

$$SNR_{dB} = 10 \log_{10} \left(\frac{P_S}{P_N} \right) = 10 \log_{10}(P_S) - 10 \log_{10}(P_N). \qquad (12)$$

Jak szum wpływa na sygnał?



Rysunek: Przykładowa obserwacja, na którą nałożono bały szum w zależności od współczynnika SNR. Źródło: Opracowanie własne

Ukryte łańcuchy Markowa - predykcja

Zajmujemy się zagadnieniem rozpoznawania izolowanych słów, które można sprowadzić do równania:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \{P(w|O)\}, \tag{13}$$

gdzie w to klasa słowa, a $O = (O_1, \ldots, O_T)$ to wektor obserwacji.
Z twierdzenia Bayes'a

$$\hat{w} = \underset{w}{\operatorname{argmax}} \left\{ \frac{P(O|w)P(w)}{P(O)} \right\} = \underset{w}{\operatorname{argmax}} \{P(O|w)P(w)\}. \tag{14}$$

Uwaga

W zagadnieniu izolowanych słów pomija się $P(w)$, ponieważ prawdopodobieństwo wystąpienia dowolnej klasy słowa powinno być takie samo dla każdej klasy.

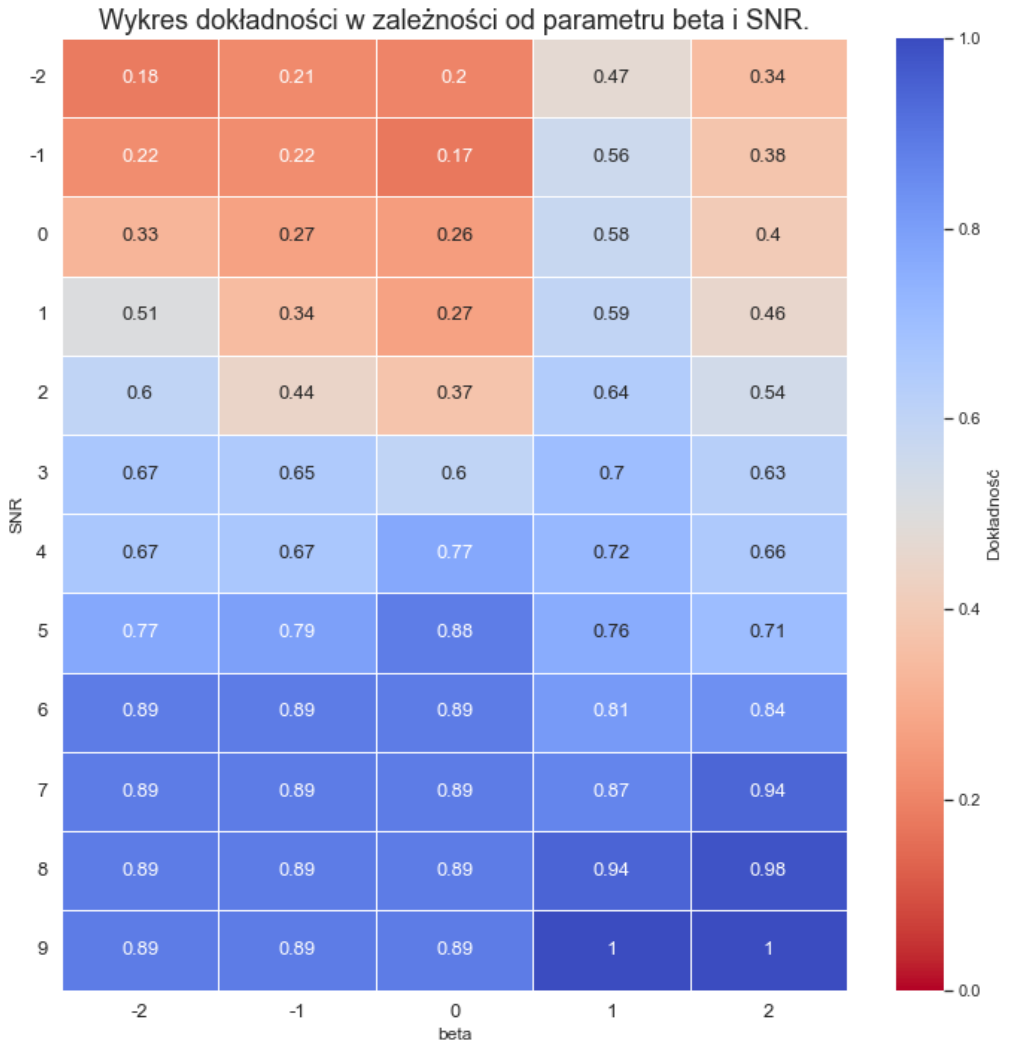
Ukryte łańcuchy Markowa - predykcja

W naszym problemie każde osobne słowo możemy przedstawić jako ukryty łańcuch Markowa. Jeśli oznaczymy taki HMM jako λ, to możemy zapisać:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \{P(O|\lambda)\}. \tag{15}$$

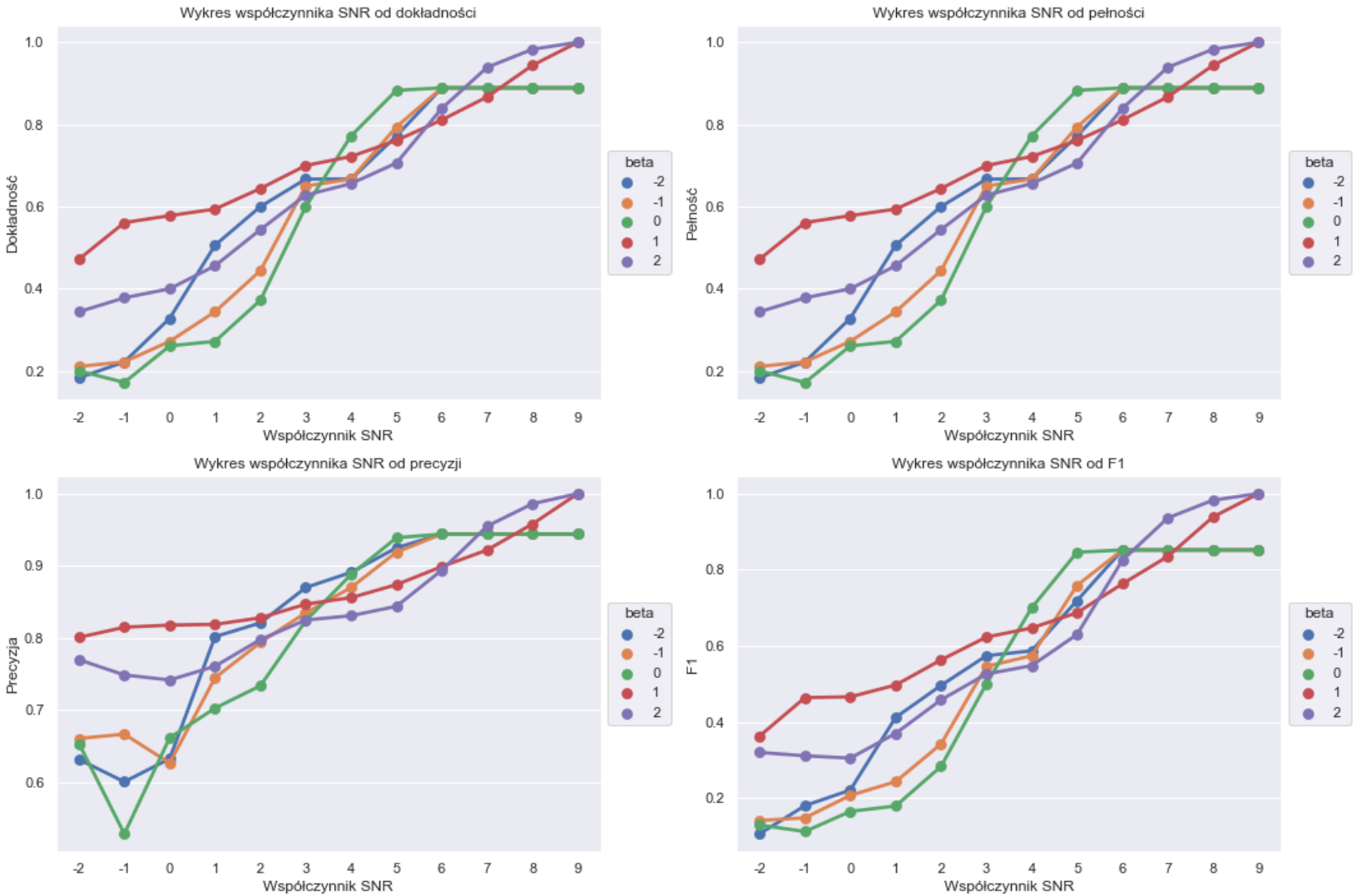
Czyli problem znajdowania słowa w zastąpiliśmy problemem wybrania najbardziej prawdopodobnego λ odpowiadającego za w.

Model bazowy - wyniki



Rysunek: Źródło: Opracowanie własne

Model bazowy - wyniki



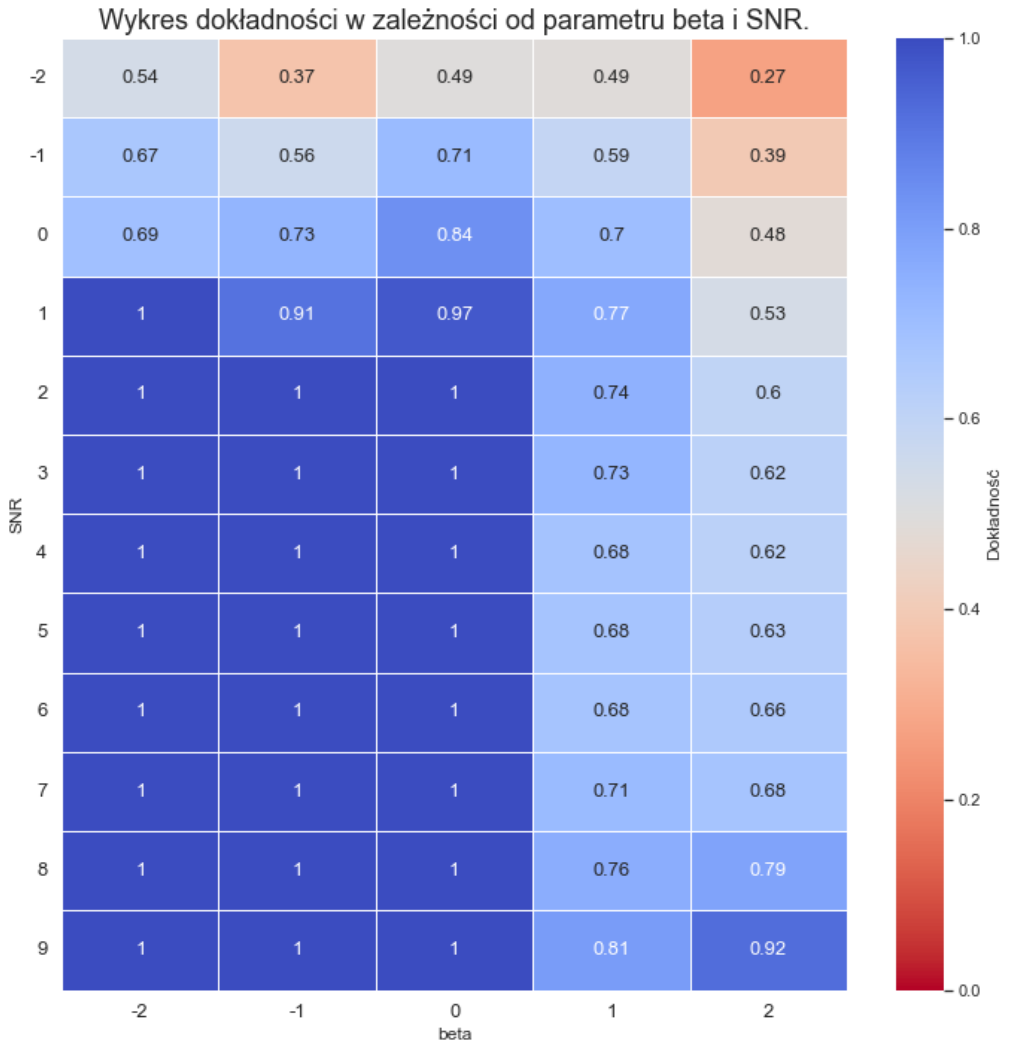
Rysunek: Źródło: Opracowanie własne

Model trenowany na zaszumionych obserwacjach

β	-2	-1	0	1	2
SNR_{dB}					
-2	0.356	0.161	0.294	0.017	-0.072
-1	0.445	0.339	0.539	0.028	0.016
0	0.366	0.461	0.578	0.122	0.083
1	0.494	0.567	0.7	0.178	0.077
2	0.4	0.556	0.628	0.1	0.056
3	0.333	0.35	0.4	0.028	-0.011
4	0.333	0.333	0.228	-0.039	-0.039
5	0.228	0.206	0.117	-0.083	-0.073
6	0.111	0.111	0.111	-0.133	-0.183
7	0.111	0.111	0.111	-0.156	-0.261
8	0.111	0.111	0.111	-0.188	-0.194
9	0.111	0.111	0.111	-0.194	-0.078

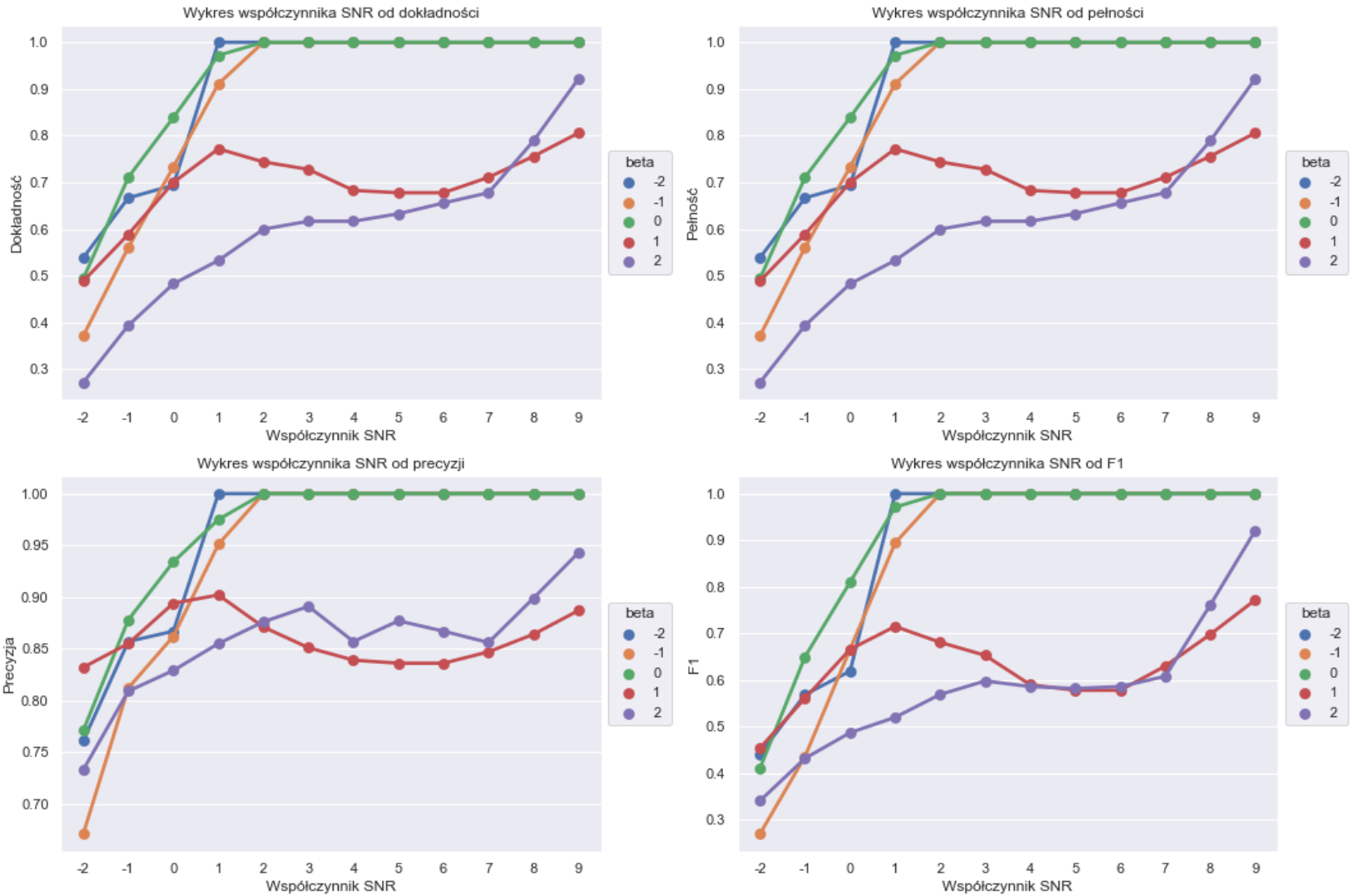
Tabela: Zwroty wyników dokładności modelu wytrenowanego na zaszumionych danych do modelu bazowego. Źródło: Opracowanie własne

Model trenowany na zaszumionych obserwacjach - wyniki



Rysunek: Źródło: Opracowanie własne

Model trenowany na zaszumionych obserwacjach - wyniki



Rysunek: Źródło: Opracowanie własne

Podsumowanie

- Ukryte Łańcuchy Markowa są przydatne w zagadnieniu rozpoznawania izolowanych słów.
- Jeśli wiemy, że w naszych obserwacjach będziemy mieli do czynienia z zakłóceniami, to warto w sposób sztuczny zaszumić zbiór treningowy.
- Przy wyborze optymalnych parametrów zaszumienia zbioru kierowaliśmy się poprawą dokładności, lecz można to uogólnić na monitorowanie dowolnej metryki.

Dziękuję za uwagę!

Literatura

[1] *Federal Standard 1037C*. Tech. rep. Institute for Telecommunication Sciences. Institute for Telecommunication Sciences, National Telecommunications and Information Administration (ITS-NTIA), 2018.

[2] Google. *Google Text-to-Speech*. https://gtts.readthedocs.io/en/latest/index.html. Dostęp: 02.01.2022.

[3] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall, 2001, pp. 314–316.

[4] D. H. Johnson. "Signal-to-noise ratio". In: *Scholarpedia* 1.12 (2006). revision #126771, p. 2088. DOI: 10.4249/scholarpedia.2088.

[5] B. H. Juang and L. R. Rabiner. "Hidden Markov Models for Speech Recognition". In: *Technometrics* 33.3 (1991), pp. 251–272.

[6] B. H. Juang L. R. Rabiner. "An introduction to hidden Markov models". In: *IEEE ASSP Mag.* vol. 3, no. 1 (1986), pp. 4–16.

[7] Petre Stoica and Randolph L. Moses. *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005.

[8] Joseph S. Wisniewski. *Colors of noise pseudo FAQ, version 1.3*. https://web.archive.org/web/20110430151608/https://www.ptpart.co.uk/colors-of-noise. Dostęp: 03.01.2022.