Czy komputery potrafią mówić?

Innowacyjne aplikacje wykorzystujące przetwarzanie dźwięku i mowy

Wydział Informatyki PB,
Katedra Mediów Cyfrowych i Grafiki Komputerowej
dr inż. Paweł Tadejko, p.tadejko@pb.edu.pl

Plan prezentacji

- Przetwarzanie mowy
- Systemy typu tekst-mowa
- Sposoby syntezy
- Po co to wszystko?
- Przykłady wykorzystania

Mowa i komputery

- Kodowanie sygnału mowy
  - w komunikacji między użytkownikami
- Rozpoznawanie i rozumienie mowy
  - przez komputer
- Poprawa jakości dźwięku / mowy
  - przez komputer
- Synteza mowy
  - przez komputer

Dźwięk odtwarzany przez komputer

Płyta muzyczna CD – jakość dźwięku

- standardowa płyta CD z muzyką dostępna w sprzedaży,
- około 15-20 utworów muzycznych
- bardzo dobra jakość muzyki stereo
- cyfrowy zapis muzyki (ang. digital)
- tzw. standard CD Audio

Pliki MP3 – jakość dźwięku

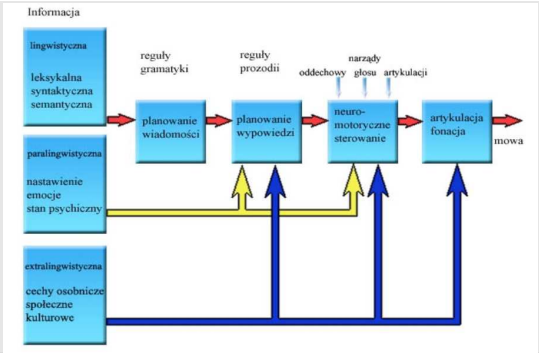
- MP3 - algorytm kompresji stratnej dźwięku
- kompresji – próbujemy „upakować” więcej informacji na mniejszej przestrzeni w komputerze
- stratnej – tracimy po drodze część informacji, która jest najmniej istotna
- możemy wybrać jakość MP3: 256, 192, 128 kbits
- dzięki temu możemy zapisać na zwykłej płycie CD o pojemności 700 MB – od 150 do 300 utworów

Dźwięk syntezowany – jakość dźwięku

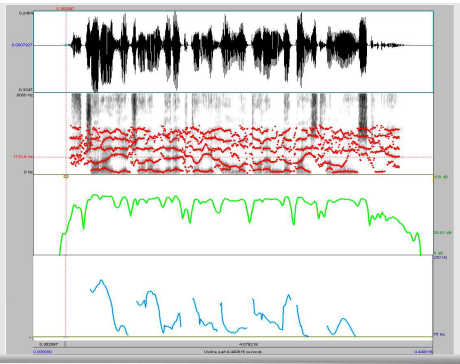
- tzw. polifoniczny (ang. chip tune)
- popularność zyskały w komputerach domowych oraz konsolach gier wideo na przełomie 1980/90
- stosowane do dziś, np. jako dzwonki polifoniczne w tel. komórkowym
- przykłady „chiptune” w serwisie YouTube
- na zwykłej płycie CD o pojemności 700 MB – od 15000 do 50000 utworów

Mowa jako sygnał cyfrowy

Jak mówi człowiek?



Jak to „słyszy” komputer?



Dlaczego chcemy, żeby komputery mówiły?

- Naturalność komunikacji:
  - Mowa jest najbardziej skutecznym (i na ogół najszybszym), łatwym i powszechnym sposobem porozumiewania się
- Skuteczność:
  - W niektórych sytuacjach jest jedynym, możliwym środkiem porozumienia się

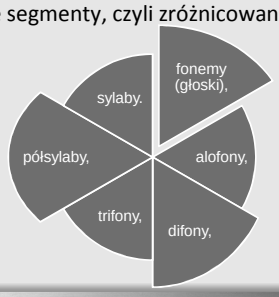
Dlaczego chcemy, żeby komputery mówiły?

- Ekspresja:
  - Pewne sytuacje, stany emocjonalne, nie są do oddania bez użycia mowy (języka naturalnego)
- Niekiedy jedyny środek komunikacji bezpośredniej:
  - Telefon, radiotelefon itp. z osobami prowadzącymi pojazdy, maszyny itp.

Synteza mowy

Analiza mowy

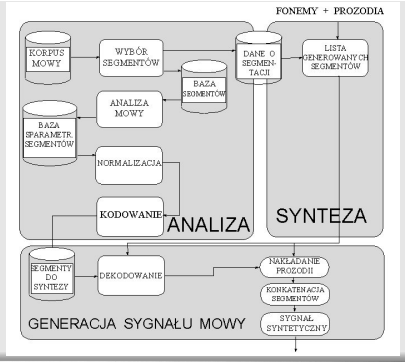
- Analiza mowy wymaga, by w sygnale będącym ciągłą sekwencją dźwięków, wyodrębnić charakterystyczne, stosunkowo niewielkie segmenty, czyli zróżnicowane jednostki akustyczne



Synteza mowy

- synteza z reguł, wykorzystująca wiedzę fonetyczno-akustyczną
- synteza słownikowa, wykorzystująca obszerne bazy danych, polegająca na łączeniu dłuższych lub krótszych elementów fonetyczno-akustycznych
- Która jest lepsza?

Schemat systemu syntezy



Podstawy informatyczne

For a Vowel, Formants are the Roots of $1+\sum a_k z^{-k}$:

$$\prod_{n=1}^{N}(1-\lambda_n z^{-1})(1-\lambda_n^* z^{-1}) = 1+\sum_{k=1}^{2N} a_k z^{-k}$$

$$\lambda_n = e^{(\pi B_n + 2\pi F_n)/F_s}$$

In the Time Domain:

$$S(z) = \frac{G}{1+\sum_{k=1}^{2N} a_k z^{-k}} \Rightarrow s[n] = G\delta[n] - \sum_{k=1}^{2N} a_k s[n-k]$$

Autocorrelation:

$$R[i] = \frac{1}{L}\sum_{n=1}^{L} s[n]s[n-i]$$

$$R[i] = \frac{1}{L}\sum_{n=1}^{L}\left(G\delta[n]s[n-i] - \sum_{k=1}^{2N} a_k s[n-k]s[n-i]\right) \approx G\bar{s}_i - \sum_{k=1}^{2N} a_k R[k-i]$$

$$\text{where } \bar{s}_i = \frac{1}{L}\sum_{n=1}^{L} s[n-i] \approx 0$$

$$\begin{bmatrix} a_1 \\ \vdots \\ a_{2N} \end{bmatrix} = \begin{bmatrix} R[0] & \dots & R[2N-1] \\ \vdots & & \vdots \\ R[2N-1] & \dots & R[0] \end{bmatrix}^{-1} \begin{bmatrix} R[1] \\ \vdots \\ R[2N] \end{bmatrix}$$

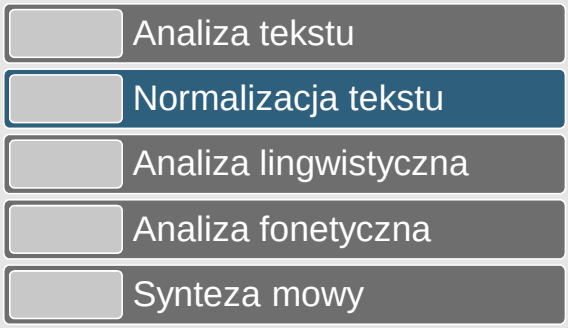
System syntezy mowy

- Analiza tekstu
- Normalizacja tekstu
- Analiza lingwistyczna
- Analiza fonetyczna
- Synteza mowy

Synteza mowy: Analiza tekstu

- Moduł analizy tekstu określa typ i strukturę przetwarzanego dokumentu,
- dokonuje konwersji nieortograficznych znaków,
- rozbioru gramatycznego,
- analizy syntaktycznej,
- leksykalnej.

System syntezy mowy



Synteza mowy: Normalizacja tekstu

- Normalizacja tekstu polega na ujednoliceniu konwersji
- wszystkich symboli, liczb i znaków nieortograficznych w transkrypcji ortograficznej,
- w postaci umożliwiającej następnie ich konwersję na ciąg znaków transkrypcji fonetycznej.

System syntezy mowy

- Analiza tekstu
- Normalizacja tekstu
- **Analiza lingwistyczna**
- Analiza fonetyczna
- Synteza mowy

Synteza mowy: Analiza lingwistyczna

- Analiza lingwistyczna tekstu obejmuje
- wybrane elementy syntaktyczne i semantyczne takie jak słowo, fraza, zdanie, wypowiedź
- by ocenić ich wpływ na samą wymowę i cechy prozodyczne

System syntezy mowy

- Analiza tekstu
- Normalizacja tekstu
- Analiza lingwistyczna
- **Analiza fonetyczna**
- Synteza mowy

Synteza mowy: Analiza fonetyczna

- Ma na celu dokonanie konwersji wyrazów
- przedstawionych w postaci kodu ortograficznego na kod fonetyczny
- z dodatkowymi informacjami (np. dotyczącymi akcentu), określającymi ich wymowę.

System syntezy mowy

- Analiza tekstu
- Normalizacja tekstu
- Analiza lingwistyczna
- Analiza fonetyczna
- Synteza mowy

Synteza mowy: Synteza mowy

- Moduł ten generuje akustyczny sygnał mowy,
- na podstawie sekwencji określonych fonemów
- uzyskanych na podstawie przetwarzania tekstu, wzorców iloczasowych, konturu melodycznego i obwiedni amplitudy

Po co to wszystko?

Przykłady zastosowań

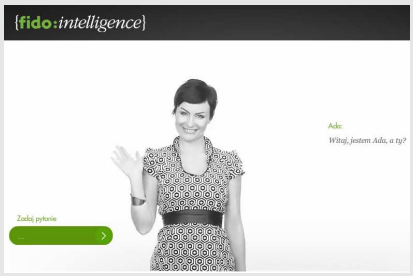
- Urządzenie do czytanie książek dla osób niewidomych
- Telefony komórkowe
- Systemy zapowiedzi słownych
- Kioski informacyjne
- Zabawki
- Urządzenia audio/wideo
- Nawigacje i innego rodzaju przewodniki

Przykłady zastosowań - Linguboty

- Lingubot (bot, chater bot) wirtualny rozmówca na stronach WWW,
- program tworzony do pełnienia zadań automatycznej i dobrze poinformowanej pomocy klientom dużych firm (banków, firm telekomunikacyjnych, ubezpieczeniowych, finansowych)

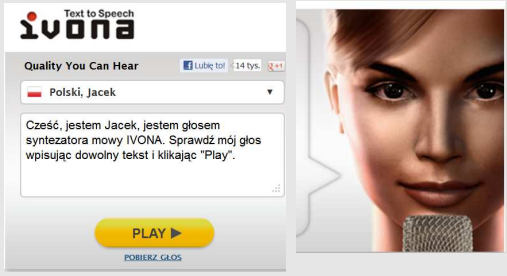
Przykłady zastosowań - Linguboty

- Polski przedstawiciel fidointeractive
  - fidointelligence.pl/pl/wirtualny-asystent

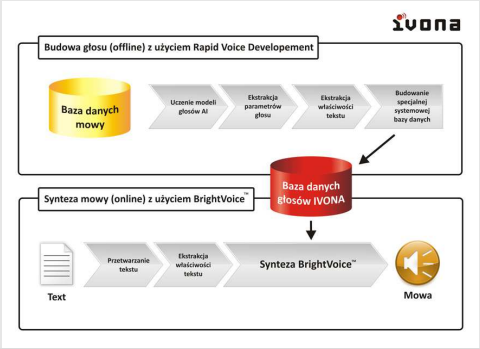


Przykłady zastosowań - IVONA

- Syntezator mowy IVONA
- http://www.ivona.com/pl/



Przykłady zastosowań - IVONA



Polskie syntezatory mowy

- SynTalk - opracowany przez firmę Neurosoft
  - Pierwszy programowy syntezer mowy (TTS) w Polsce (1994)
- UNIT-SELECTION –Korpusowy Syntezator mowy (Polsko Japońska Wyższa Szkoła Technik Komputerowych)
  - http://syntezamowy.pjwstk.edu.pl/korpus.html
- Syntezator mowy IVONA
  - http://www.ivona.com/pl/
- Milena – syntezator mowy polskiej dla środowiska Linux
  - http://milena.polip.com/

Podsumowanie

- Czy może syntezator coś zaśpiewać?

Dziękuję za uwagę

Pytania?

Czy komputery potrafią mówić?

Innowacyjne aplikacje wykorzystujące przetwarzanie dźwięku i mowy

Wydział Informatyki PB,

Katedra Mediów Cyfrowych i Grafiki Komputerowej

dr inż. Paweł Tadejko, p.tadejko@pb.edu.pl

Plan prezentacji

- Przetwarzanie mowy
- Systemy typu tekst-mowa
- Sposoby syntezy
- Po co to wszystko?
- Przykłady wykorzystania

Mowa i komputery

- Kodowanie sygnału mowy
  - w komunikacji między użytkownikami
- Rozpoznawanie i rozumienie mowy
  - przez komputer
- Poprawa jakości dźwięku / mowy
  - przez komputer
- Synteza mowy
  - przez komputer

Dźwięk odtwarzany przez komputer

Płyta muzyczna CD – jakość dźwięku

- standardowa płyta CD z muzyką dostępna w sprzedaży,
- około 15-20 utworów muzycznych
- bardzo dobra jakość muzyki stereo
- cyfrowy zapis muzyki (ang. digital)
- tzw. standard CD Audio

Pliki MP3 – jakość dźwięku

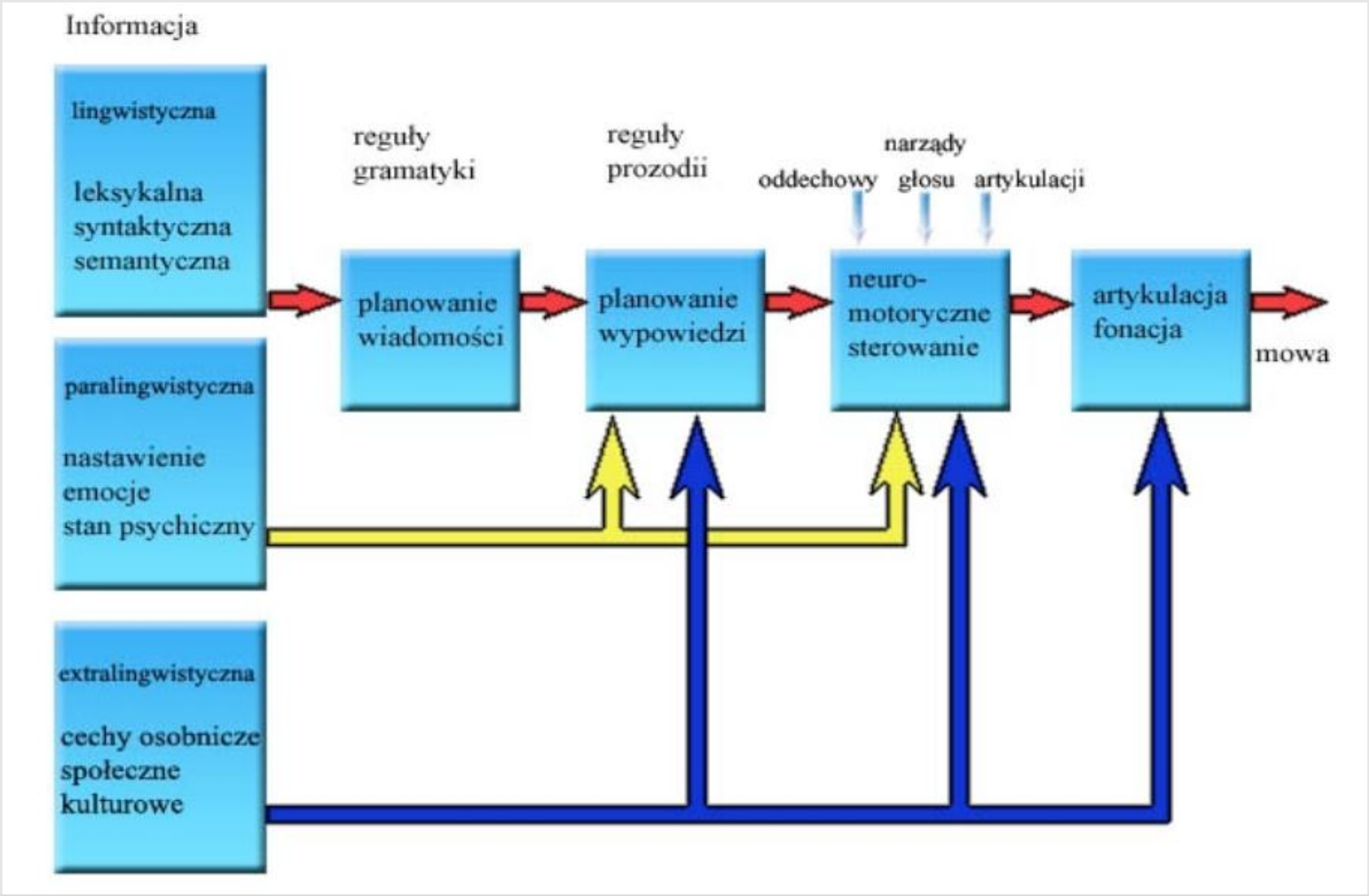
- MP3 - algorytm kompresji stratnej dźwięku
- kompresji – próbujemy „upakować” więcej informacji na mniejszej przestrzeni w komputerze
- stratnej – tracimy po drodze część informacji, która jest najmniej istotna
- możemy wybrać jakość MP3: 256, 192, 128 kbits
- dzięki temu możemy zapisać na zwykłej płycie CD o pojemności 700 MB – od 150 do 300 utworów

Dźwięk syntezowany – jakość dźwięku

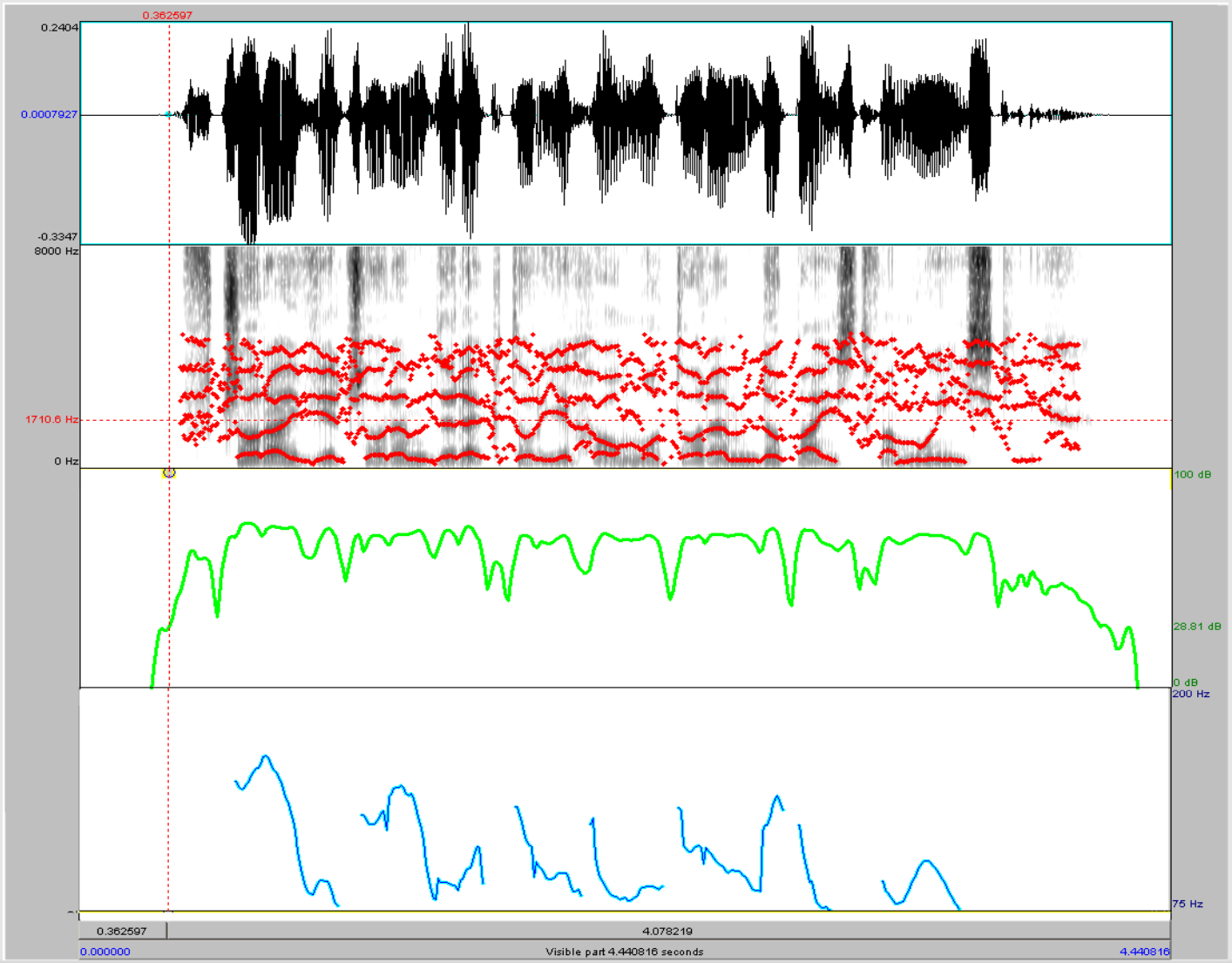
- tzw. polifoniczny (ang. chip tune)
- popularność zyskały w komputerach domowych oraz konsolach gier wideo na przełomie 1980/90
- stosowane do dziś, np. jako dzwonki polifoniczne w tel. komórkowym
- przykłady „chiptune” w serwisie YouTube
- na zwykłej płycie CD o pojemności 700 MB – od 15000 do 50000 utworów

Mowa jako sygnał cyfrowy

Jak mówi człowiek?



Jak to „słyszy” komputer?



Dlaczego chcemy, żeby komputery mówiły?

- Naturalność komunikacji:
  - Mowa jest najbardziej skutecznym (i na ogół najszybszym), łatwym i powszechnym sposobem porozumiewania się
- Skuteczność:
  - W niektórych sytuacjach jest jedynym, możliwym środkiem porozumienia się

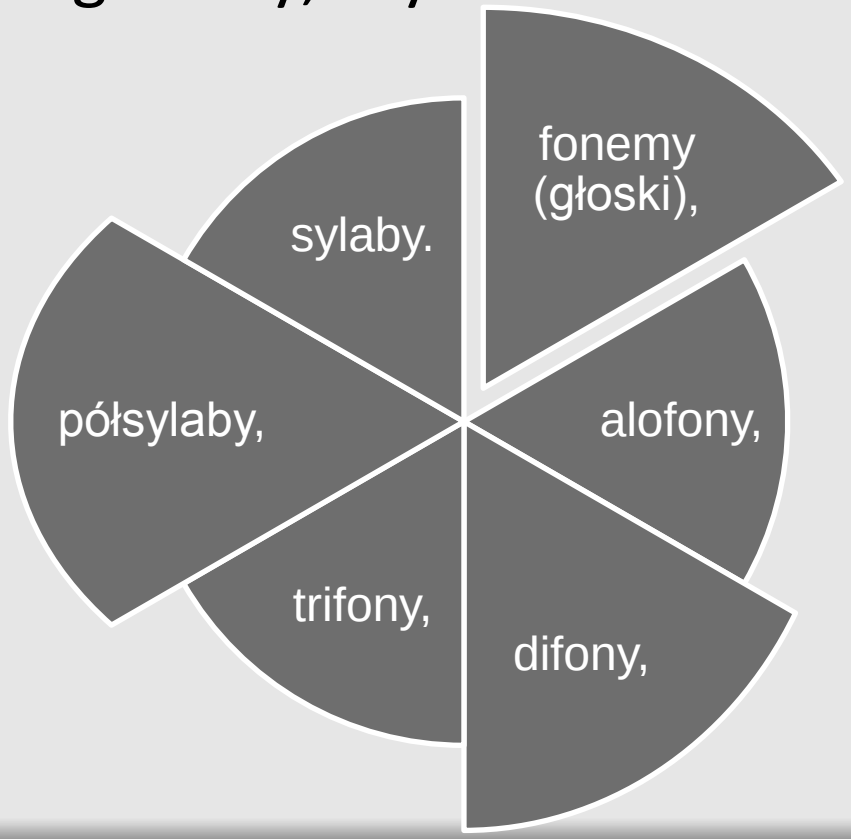
Dlaczego chcemy, żeby komputery mówiły?

- Ekspresja:
  - Pewne sytuacje, stany emocjonalne, nie są do oddania bez użycia mowy (języka naturalnego)
- Niekiedy jedyny środek komunikacji bezpośredniej:
  - Telefon, radiotelefon itp. z osobami prowadzącymi pojazdy, maszyny itp.

Synteza mowy

Analiza mowy

- Analiza mowy wymaga, by w sygnale będącym ciągłą sekwencją dźwięków, wyodrębnić charakterystyczne, stosunkowo niewielkie segmenty, czyli zróżnicowane jednostki akustyczne

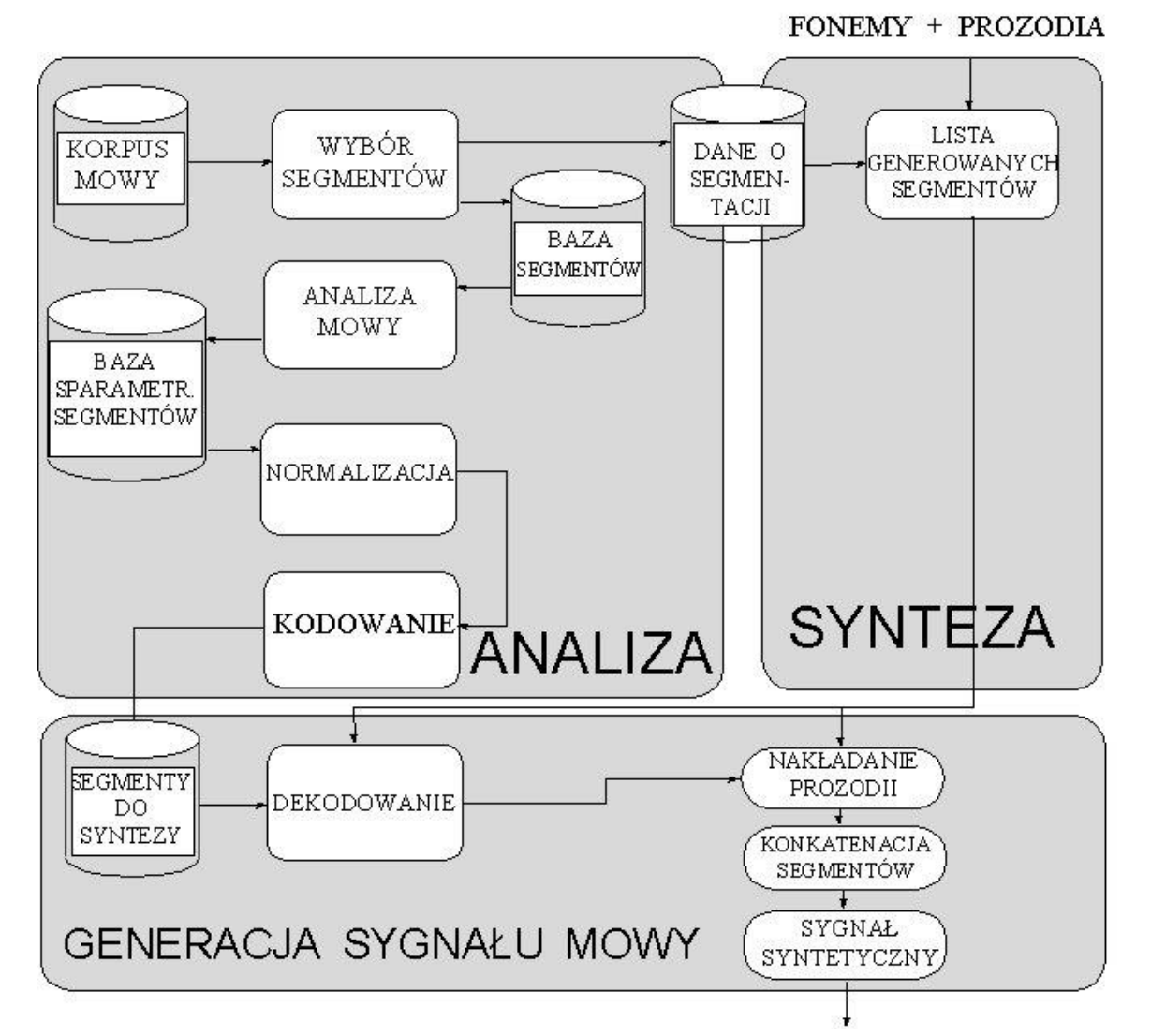


Synteza mowy

- synteza z reguł, wykorzystująca wiedzę fonetyczno-akustyczną
- synteza słownikowa, wykorzystująca obszerne bazy danych, polegająca na łączeniu dłuższych lub krótszych elementów fonetyczno-akustycznych

- Która jest lepsza?

Schemat systemu syntezy



Podstawy informatyczne

For a Vowel, Formants are the Roots of $1 + \sum a_k z^{-k}$:

$$\prod_{n=1}^{N}(1-\lambda_n z^{-1})(1-\lambda_n^* z^{-1}) = 1 + \sum_{k=1}^{2N} a_k z^{-k}$$

$$\lambda_n = e^{(\pi B_n + 2\pi F_n)/F_s}$$

In the Time Domain:

$$S(z) = \frac{G}{1+\sum_{k=1}^{2N} a_k z^{-k}} \Rightarrow s[n] = G\delta[n] - \sum_{k=1}^{2N} a_k s[n-k]$$

Autocorrelation:

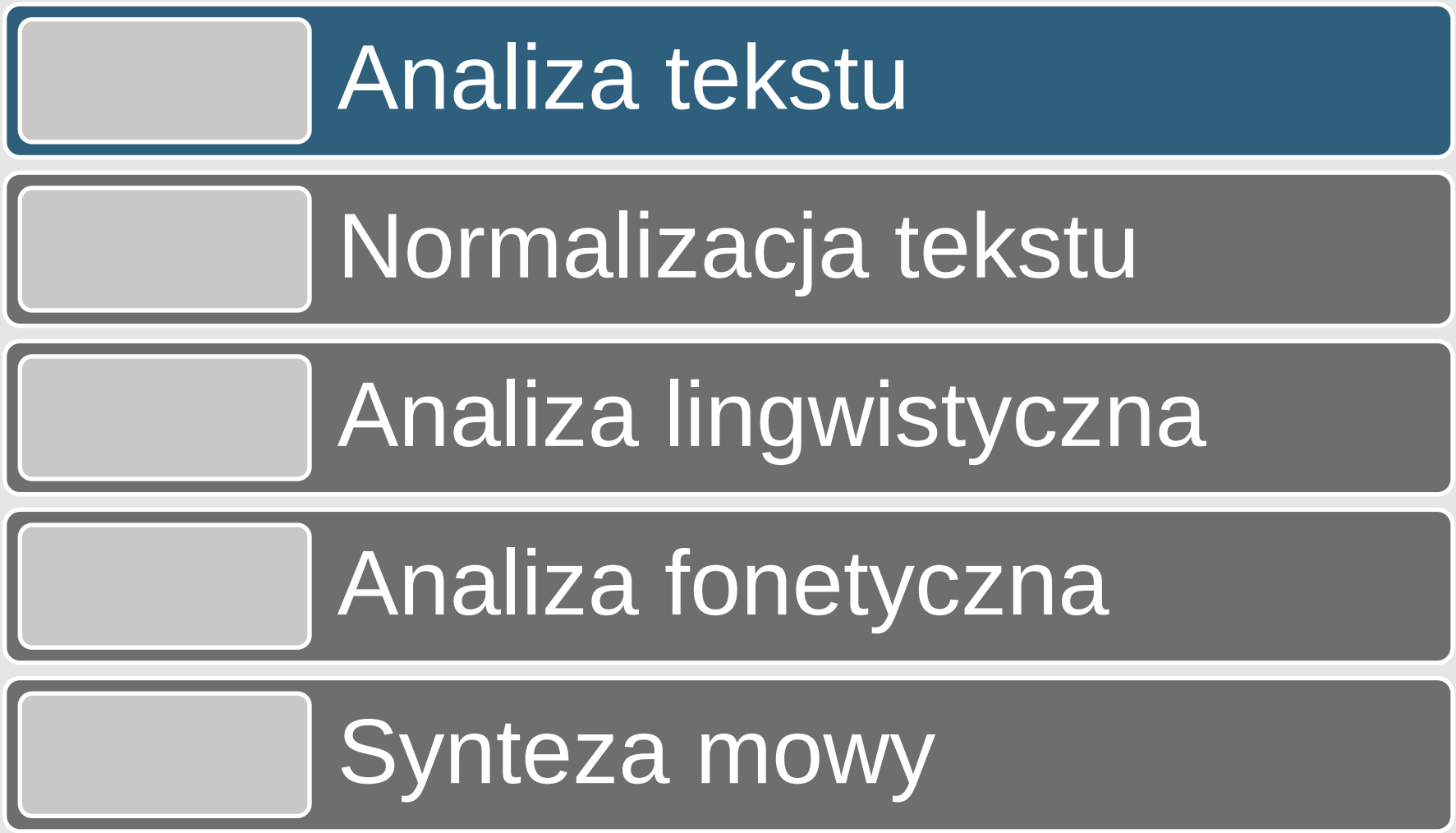
$$R[i] = \frac{1}{L}\sum_{n=1}^{L} s[n]s[n-i]$$

$$R[i] = \frac{1}{L}\sum_{n=1}^{L}\left(G\delta[n]s[n-i] - \sum_{k=1}^{2N} a_k s[n-k]s[n-i]\right) \approx G\bar{s}_i - \sum_{k=1}^{2N} a_k R[k-i]$$

$$\text{where } \bar{s}_i = \frac{1}{L}\sum_{n=1}^{L} s[n-i] \approx 0$$

$$\begin{bmatrix} a_1 \\ \vdots \\ a_{2N} \end{bmatrix} = \begin{bmatrix} R[0] & \dots & R[2N-1] \\ \vdots & & \vdots \\ R[2N-1] & \dots & R[0] \end{bmatrix}^{-1} \begin{bmatrix} R[1] \\ \vdots \\ R[2N] \end{bmatrix}$$

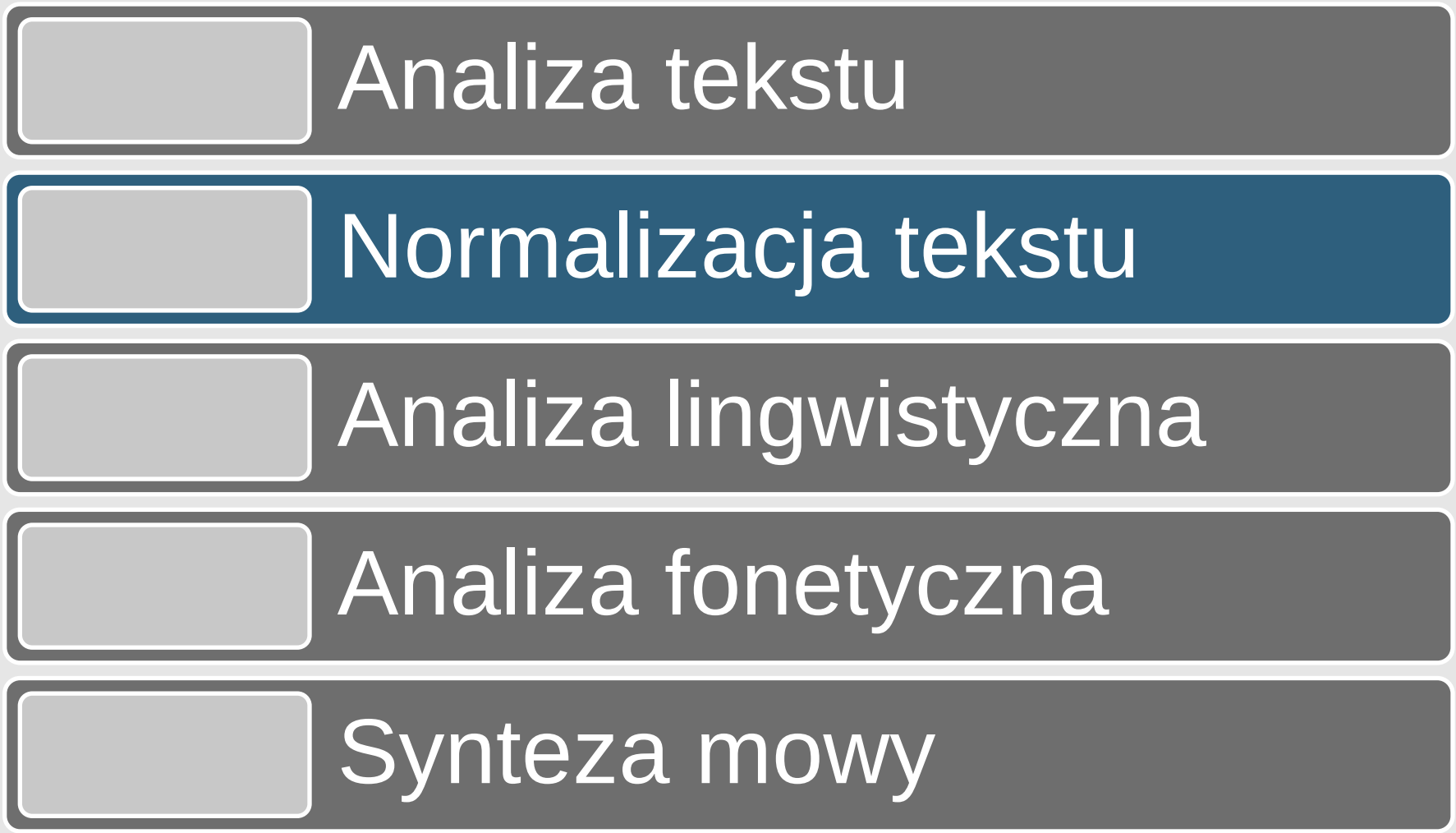
System syntezy mowy



Synteza mowy: Analiza tekstu

- Moduł analizy tekstu określa typ i strukturę przetwarzanego dokumentu,
- dokonuje konwersji nieortograficznych znaków,
- rozbioru gramatycznego,
- analizy syntaktycznej,
- leksykalnej.

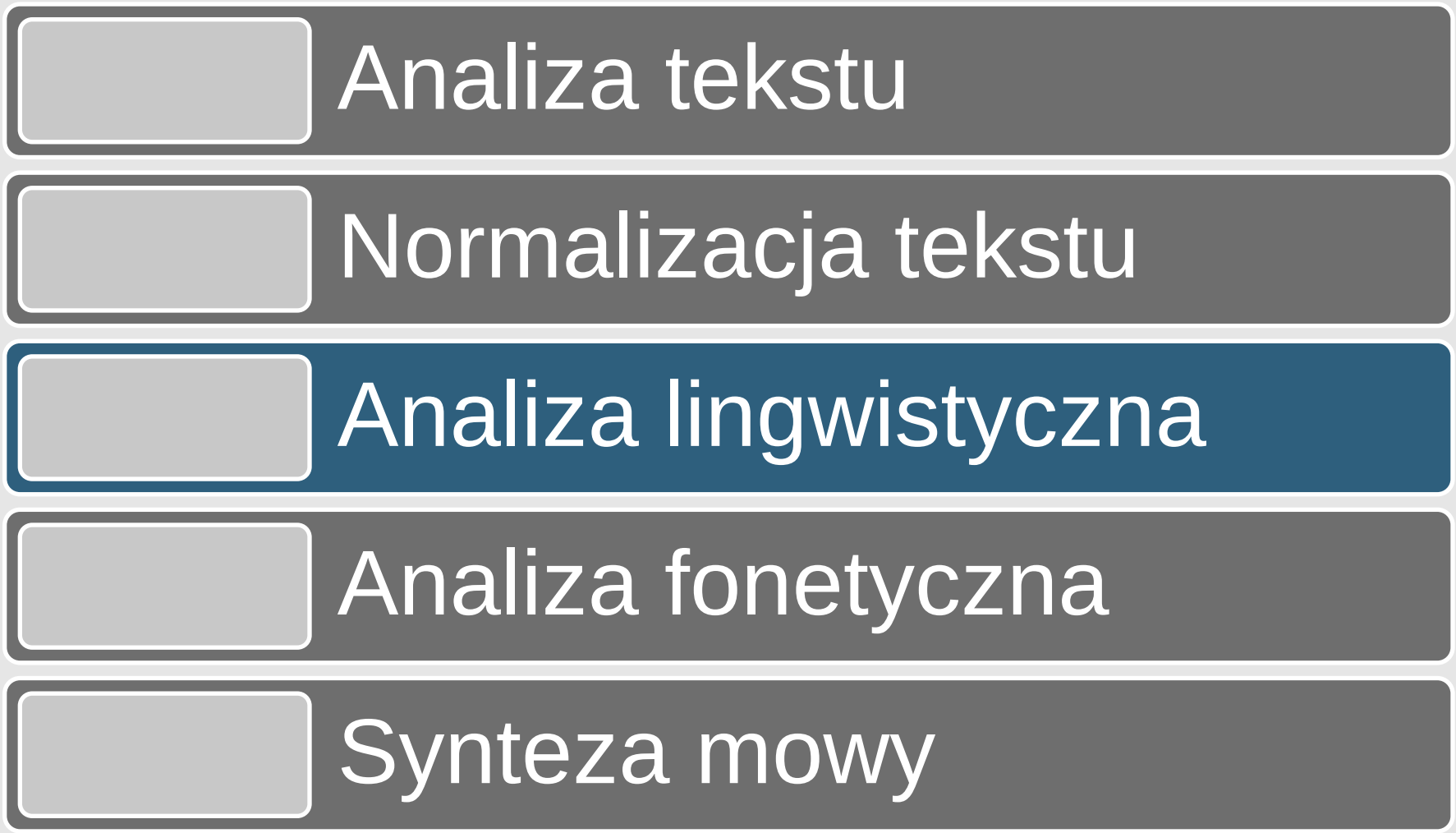
System syntezy mowy



Synteza mowy: Normalizacja tekstu

- Normalizacja tekstu polega na ujednoliceniu konwersji
- wszystkich symboli, liczb i znaków nieortograficznych w transkrypcji ortograficznej,
- w postaci umożliwiającej następnie ich konwersję na ciąg znaków transkrypcji fonetycznej.

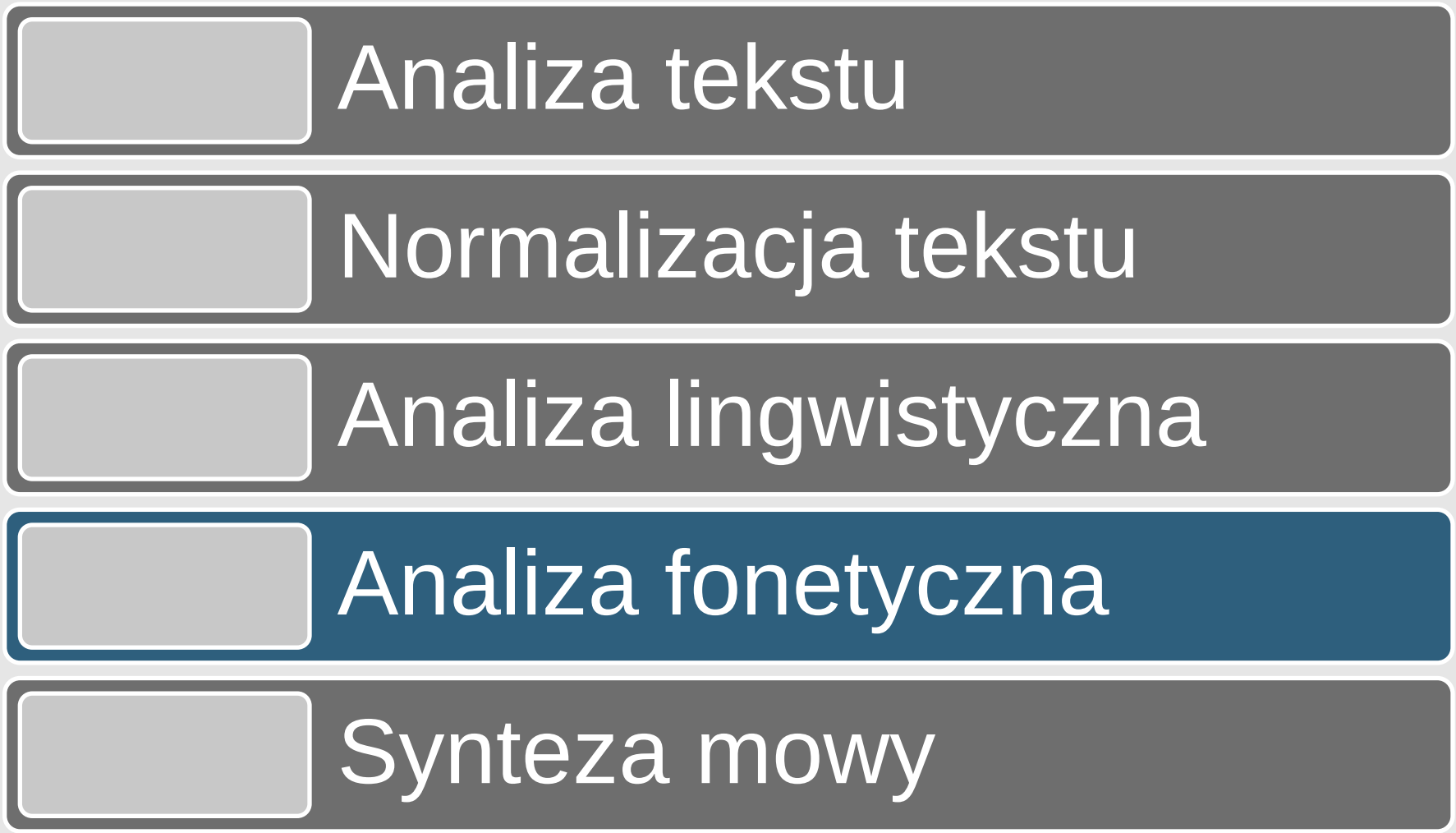
System syntezy mowy



Synteza mowy: Analiza lingwistyczna

- Analiza lingwistyczna tekstu obejmuje
- wybrane elementy syntaktyczne i semantyczne takie jak słowo, fraza, zdanie, wypowiedź
- by ocenić ich wpływ na samą wymowę i cechy prozodyczne

System syntezy mowy



Synteza mowy: Analiza fonetyczna

- Ma na celu dokonanie konwersji wyrazów
- przedstawionych w postaci kodu ortograficznego na kod fonetyczny
- z dodatkowymi informacjami (np. dotyczącymi akcentu), określającymi ich wymowę.

System syntezy mowy

- Analiza tekstu
- Normalizacja tekstu
- Analiza lingwistyczna
- Analiza fonetyczna
- Synteza mowy

Synteza mowy: Synteza mowy

- Moduł ten generuje akustyczny sygnał mowy,
- na podstawie sekwencji określonych fonemów
- uzyskanych na podstawie przetwarzania tekstu, wzorców iloczasowych, konturu melodycznego i obwiedni amplitudy

Po co to wszystko?

Przykłady zastosowań

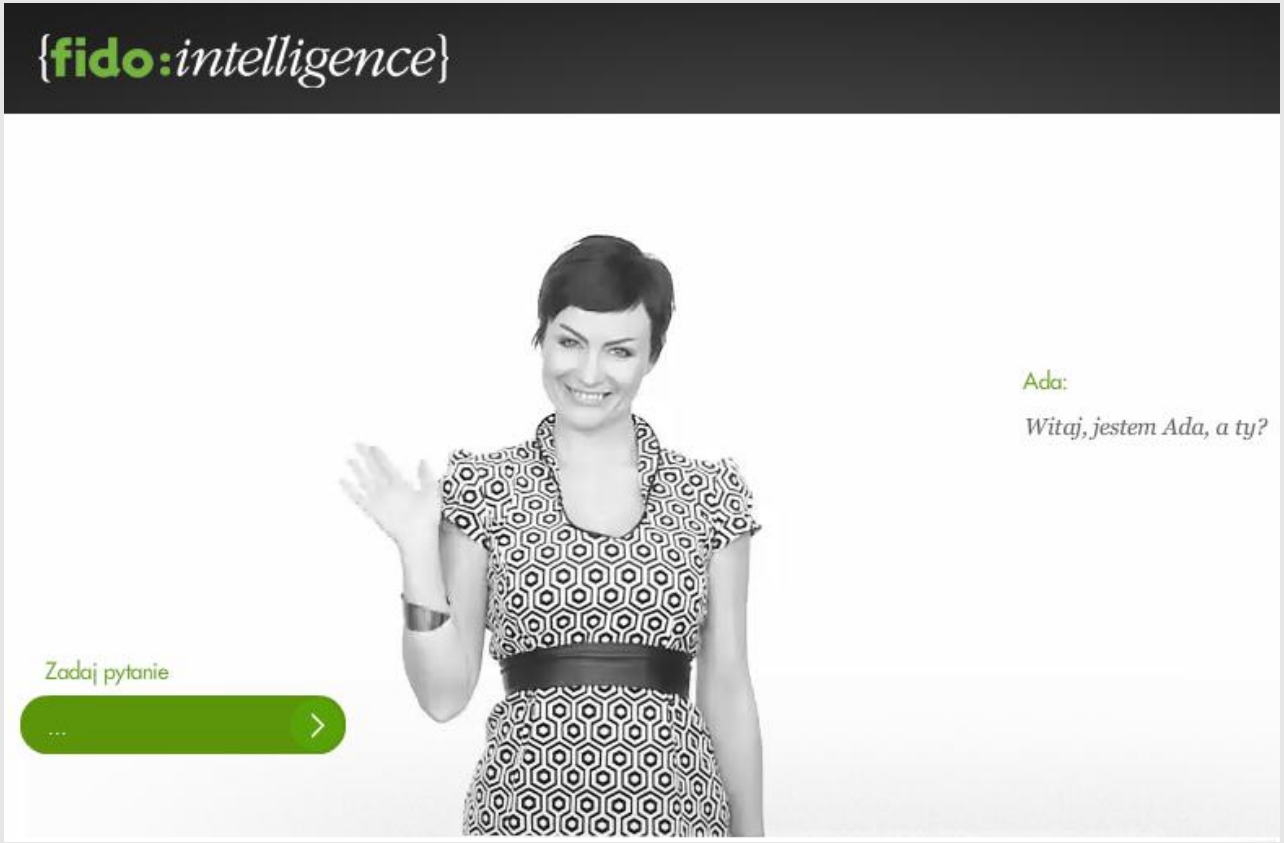
- Urządzenie do czytanie książek dla osób niewidomych
- Telefony komórkowe
- Systemy zapowiedzi słownych
- Kioski informacyjne
- Zabawki
- Urządzenia audio/wideo
- Nawigacje i innego rodzaju przewodniki

Przykłady zastosowań - Linguboty

- Lingubot (bot, chater bot) wirtualny rozmówca na stronach WWW,
- program tworzony do pełnienia zadań automatycznej i dobrze poinformowanej pomocy klientom dużych firm (banków, firm telekomunikacyjnych, ubezpieczeniowych, finansowych)

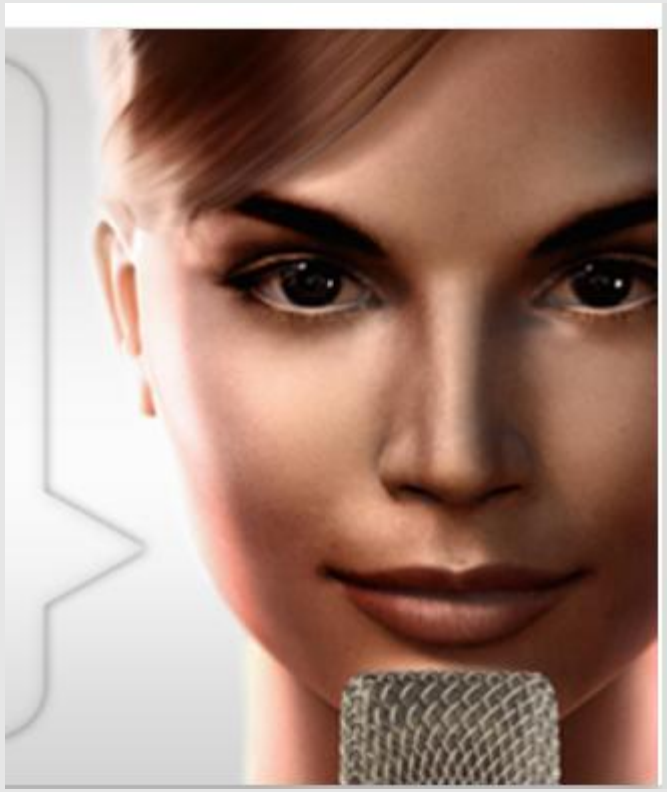
Przykłady zastosowań - Linguboty

- Polski przedstawiciel fidointeractive
  - fidointelligence.pl/pl/wirtualny-asystent

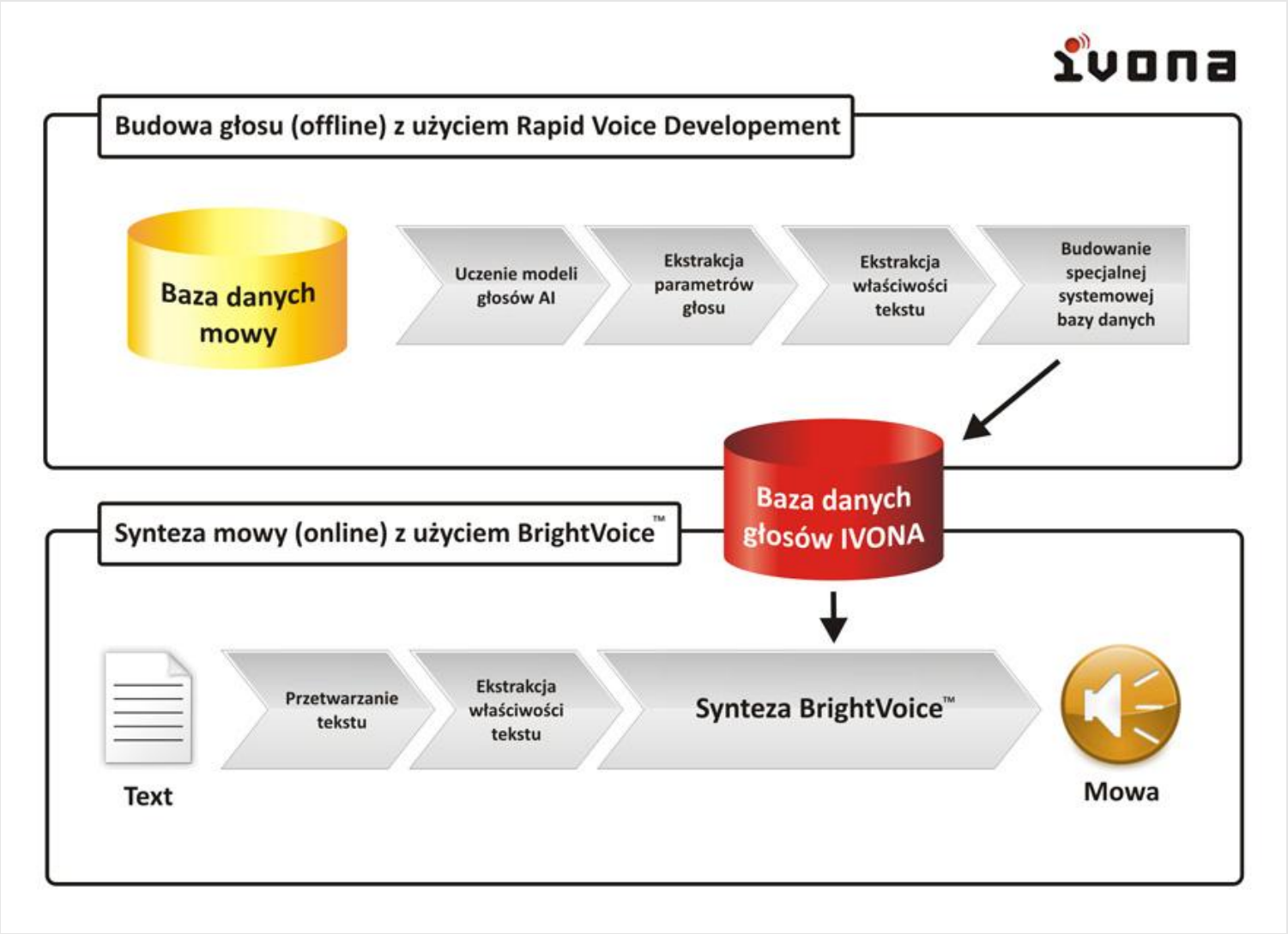


Przykłady zastosowań - IVONA

- Syntezator mowy IVONA
- http://www.ivona.com/pl/



Przykłady zastosowań - IVONA



Polskie syntezatory mowy

- SynTalk - opracowany przez firmę Neurosoft
  - Pierwszy programowy syntezer mowy (TTS) w Polsce (1994)
- UNIT-SELECTION –Korpusowy Syntezator mowy (Polsko Japońska Wyższa Szkoła Technik Komputerowych)
  - http://syntezamowy.pjwstk.edu.pl/korpus.html
- Syntezator mowy IVONA
  - http://www.ivona.com/pl/
- Milena – syntezator mowy polskiej dla środowiska Linux
  - http://milena.polip.com/

Podsumowanie

- Czy może syntezator coś zaśpiewać?

Dziękuję za uwagę

Pytania?