Synteza mowy (TTS)

Tomasz Hołda
Krzysztof Gonet

Agenda

- Czym jest synteza mowy?
- Krótka historia syntezy mowy
- Rodzaje syntezy mowy
- Proces syntezy mowy
- Metoda konktenatywna
- Metoda formantu
- Metoda artykulacyjna

Synteza mowy (TTS)

Synteza mowy - to dział przetwarzania mowy polegający na mechanicznej zamianie tekstu zapisanego w postaci znakowej na wypowiedź (mowę) w postaci dźwiękowej.

Syntezatora mowy - maszyna lub program komputerowy zamieniający tekst na mowę.

Historia syntezy mowy

1769: *Wolfgang von Kempelen* - jedna z pierwszych mechanicznych maszyn mówiących.

Lata 40 XX w.: *Frank Cooper* - system odtwarzania wzorców dźwięku oparty o widmo częstotliwości.

1984: *Apple* - wbudowany syntezator mowy MacInTalk w komputerach Macintosh.

2011: *Apple* - Siri, pierwszy „inteligentny agent” w telefonach.

2016: *Google* - Asystenta Google, odpowiedź na Cortanę i Siri.

Główne rodzaje syntezy mowy

- Synteza konkatenacji
- Synteza wyboru jednostek
- Synteza difonowa
- Synteza specyficzna dla domeny
- Synteza formantów
- Synteza artykulacyjna
- Synteza oparta na HMM (Ukryty Model Markova)
- Synteza fali sinusoidalnej
- Synteza oparta na głębokim uczeniu

Proces syntezy mowy - normalizacja

Przetwarzanie wstępne (normalizacja) - zmniejszenie niejednoznaczności, zawężenie wielu różnych sposobów, w jakie można odczytać fragment tekstu.

Przykłady:

- Zamiana dat, liczb, godzin, skrótów itd. na słowa:

  Przykładowo liczba 1843 może odnosić się do ilości rzeczy (tysiąc osiemset czterdzieści trzy) lub roku (tysiąc osiemset czterdziesty trzeci).

- Przetworzenie homografów:

  “I read the book” - wymowa “read” lub “red”.

Fonemy

Fonem – podstawowa jednostka struktury fonologicznej mowy. Jest różnorodnie definiowany, ale zwykle o jego wyróżnianiu ma decydować rozróżnianie dzięki niemu znaczenia wyrazów.

Fonem jest pojęciem abstrakcyjnym, realizowanym w rzeczywistej mowie przez głoski.

Przykład Fonemu

Przykładem fonemu jest spółgłoska szczelinowa wargowo-zębowa bezdźwięczna [Numer w międzynarodowym alfabecie fonetycznym - 128]. Podczas wymawiania fonemu:

- Modulowany jest prąd powietrza wydychanego z płuc, czyli artykulacja tej spółgłoski wymaga inicjacji płucnej i egresji.
- Tylna część podniebienia miękkiego na skutek działania mięśni zamyka dostęp do jamy nosowej, prąd powietrza uchodzi przez jamę ustną
- Prąd powietrza w jamie ustnej przepływa ponad całym językiem lub też co najmniej powietrze uchodzi wzdłuż środkowej linii języka

Polski – ‚**f**utro’ [ˈfutr̪ɔ]

Angielski – ‚**th**ink’ [fɪŋk] lub ‚**f**ootball’ [futb̪ol]

Zamiana słów na fonemy - słownik

Jedno zdanie można odczytać na wiele różnych sposobów, w zależności od znaczenia tekstu, osoby mówiącej i emocji, które chce przekazać (w lingwistyce ta idea nazywana jest prozodią i jest najtrudniejszych problemów, z którymi muszą się uporać syntezatory mowy).

Słownik fonemów musi zawierać wszystkie występujące w tekście słowa.

W praktyce jest to stosunkowo trudne zadanie.

Zamiana słów na fonemy - grafemy

Grafem – najmniejsza jednostka pisma, która często odpowiada fonemowi. Czasem jeden fonem może mieć kilka odpowiadających mu grafemów (w jęz. polskim np. rz i ż, u i ó). W alfabetach grafem najczęściej jest literą lub znakiem interpunkcyjnym

Przy użyciu grafemów, komputer może podjąć rozsądną próbę odczytania dowolnego słowa, niezależnie od tego, czy jest to prawdziwe słowo przechowywane w słowniku, obce słowo, czy też nietypowa nazwa lub termin techniczny.

Języki takie jak angielski mają dużą liczbę nieregularnych słów, które są wymawiane w bardzo różny sposób od ich zapisu - słowo “colonel” (pułkownik) wymawia się jako “kernel” a nie “coll-o-nell”

Zamiana fonemów do dźwięków

Po przekonwertowaniu tekstu (sekwencja pisanych słów) na listę fonemów (sekwencję dźwięków, które wymagają mówienia), musimy znaleźć sposób na pozyskanie tych właśnie fonemów które komputer odczytuje na głos, zamieniając tekst na mowę.

Istnieją trzy różne podejścia pozyskania fonemów:

- Metoda konkatenatywna
- Metoda formantu
- Metoda artykulacyjna

Metoda konkatenatywna

Główną cechą metody konkatenatywnej jest wykorzystywanie nagranych ludzkich głosów, które można przearanżować. Jeżeli istnieje wystarczająca liczba próbek mowy, komputer może zmienić kolejność fonemów na wiele różnych sposobów, aby stworzyć zupełnie nowe słowa i zdania.

Przykładowy zbiór ludzkich głosów - https://openslr.org/

Metoda konkatenatywna - wady i zalety

+ Jest najbardziej naturalnie brzmiącym rodzajem syntezy mowy.

- Stosowana wszędzie tam gdzie ilość tekstu do wypowiedzenia nie jest duża np. korporacyjne centrale telefoniczne (Call Center).

Metoda formantu

Opiera się na fakcie, że mowa to po prostu wzorzec dźwięku, który różni się wysokością (częstotliwością) i głośnością (amplitudą). Metoda działa tak jak syntezatory muzyczne.

Ten rodzaj syntezy mowy jest znany jako formant, ponieważ formanty są 3–5 kluczowymi (rezonansowymi) częstotliwościami dźwięku, które ludzki aparat głosowy generuje i łączy, tworząc dźwięk mowy lub śpiewu.

Metoda formantu - wady i zalety

+ Syntezatory oparte o tę metodę mogą powiedzieć absolutnie wszystko — nawet słowa, które nie istnieją lub obce słowa, których nigdy nie spotkały.
+ Syntezatory formantu są dobrym wyborem dla nawigacji GPS.
+ Mogą łatwo przełączyć się z męskiego na żeński głos (poprzez mniej więcej podwojenie częstotliwości) lub na głos dziecka (poprzez potrojenie) i potrafią mówić w dowolnym języku.

- Brzmią stosunkowo sztucznie i robotycznie.

Metoda artykulacyjna

Polega na syntezie poprzez modelowanie niezwykle skomplikowanego aparatu głosowego człowieka. Metoda eksperymentalna.

Najbardziej skomplikowaną formą syntezy artykulacyjnej byłoby skonstruowanie robota z „gadającą głową” z ruchomymi ustami, który wytwarza dźwięk w podobny sposób jak człowiek, łącząc w razie potrzeby elementy mechaniczne, elektryczne i elektroniczne.

Metoda artykulacyjna- wady i zalety

+ Teoretycznie można uzyskać najbardziej realistyczny i ludzki głos ze wszystkich trzech metod.

- Nadal zdecydowanie najmniej zbadana metoda, głównie ze względu na jej złożoność.

Dziękujemy za uwagę