© Wydawnictwo UR 2018 „Edukacja – Technika – Informatyka” nr 1/23/2018
ISSN 2080-9069 www.eti.rzeszow.pl
ISSN 2450-9221 online DOI: 10.15584/eti.2018.1.4

ARTUR ZACNIEWSKI[1], MARCIN KLEINSZMIDT[2], RADOSŁAW ZDUNEK[3]

Badania szybkości i skuteczności metod syntezy mowy na potrzeby zastosowania w urządzeniu przenośnym

Research of Speed and Accuracy of Speech Synthesis Methods for the Purposes of Deployment in Embeddeddevice

[1] Doktor inżynier, Akademia Marynarki Wojennej w Gdyni, Wydział Nawigacji i Uzbrojenia Okrętowego, Zakład Informatyki, Polska
[2] Magister inżynier, Toucan Systems, Sp. z o.o., Gdańsk, Polska
[3] Magister inżynier, Toucan Systems, Sp. z o.o., Gdańsk, Polska

Streszczenie

W artykule przeanalizowano szereg metod dotyczących syntezy mowy, mając na uwadze ich wykorzystanie w urządzeniu przenośnym. Badania realizowano na urządzeniach o zróżnicowanych parametrach, a badanymi kryteriami były skuteczność danej metody i jej szybkość. Badania są częścią projektu ToucanEye – urządzenia przenośnego z systemem sztucznej inteligencji, mającego wspomóc osoby z dysfunkcją wzroku. Pokazano również, jak ważne jest zoptymalizowanie zastosowanych metod w fazie projektu inżynierskiego w celu zapewnienia lepszej jakości pracy urządzenia i komfortu użytkownika końcowego.

Słowa kluczowe: synteza mowy, wspomaganie osób z dysfunkcją wzroku, ToucanEye

Abstract

In the article the methods concerning speech synthesis were analysed, having in mind their usage in embedded device. Research was carried out on devices with mixed parameters, and the criteria were accuracy and speed of given method. The research are part the Toucan Eye project – embedded device with artificial intelligence system able to help people with impaired sight. It was shown how important is optimization of applied methods in the phase of engineer project to ensure better quality of working device and comfort of end-user.

Keywords: speech synthesis, assisting persons with impaired sight, Toucan Eye

Wstęp

Częścią składową systemów sztucznej inteligencji wspomagających osoby z dysfunkcją wzroku bardzo często jest system syntezy mowy. Synteza mowy

jest wieloetapowym procesem mechanicznej zamiany tekstu w postaci znakowej na sygnał audio (w tym przypadku mowę), najlepiej z takim efektem, jaki dałoby przeczytanie tego tekstu przez człowieka (Tadeusiewicz, 1998, s. 194–212). W procesie syntezy wyróżnić można dwa etapy:

– przetwarzanie tekstu naturalnego NLP (ang. *Natural Language Processing*), który zawiera szereg procesów wykonywanych po stronie tekstowej, czyli przede wszystkim normalizacja tekstu, ale także zwrócenie uwagi na niuanse kontekstowe, które mogą zmieniać pożądany sposób wymówienia danego słowa,

– faza syntezy właściwej, gdzie algorytm pracujący najczęściej na odpowiednio przygotowanym zestawie próbek zajmuje się przetwarzaniem tekstu na sygnał audio (Łopatka, Czyżewski, 2010, s. 105–106).

Na urządzeniach o różnych parametrach przeprowadzono badania szybkości i skuteczności metod syntezy mowy, po to by ocenić szanse ich realizacji w czasie rzeczywistym na urządzeniu przenośnym. Użyto trzech urządzeń:

– urządzenie 1: Windows 10 Pro x64, Intel Core i5-4460, 16GB RAM,

– urządzenie 2: Windows 10 Pro x64, AMD A8-7600 Radeon R7, 4GB RAM,

– urządzenie 3: Linux Debian, Raspberry PI3, 4x ARM Cortex A53, 1 GB RAM.

Wybór odpowiedniego urządzenia oraz optymalizacja ww. metod są niezmiernie istotne dla uzyskania satysfakcjonujących wyników, stąd wyłania się ważna rola kształcenia inżynierskiego twórców i osób implementujących te metody. Dla tego konkretnego przypadku urządzenie ma być używane przez osoby niewidome i niedowidzące i to właśnie od właściwej inżynierskiej implementacji i optymalizacji konkretnych metod zależeć będzie jakość i komfort użytkowania urządzenia ToucanEye. Należy stąd wnioskować, że przydatność ww. metod w kształceniu inżynierskim wydaje się oczywista. Ich należyte praktyczne wykorzystanie przez inżyniera może pozwolić na ułatwienia kształcenia osobom z różnego rodzaju dysfunkcjami wzroku.

Metody normalizacji treści

Proces normalizacji treści zawiera mechanizmy analizy morfologicznej, syntaktycznej oraz semantycznej. Pierwszym krokiem jest normalizacja tekstu, czyli zamiana znaków niebędących literami na odpowiadającą im informację słowną. Oprócz znaków, które nie są literami, algorytmy wyszukują i zamieniają także skróty i skrótowce, a także wyrazy zaczerpnięte z innych języków. Normalizacja pozwala także na odrzucenie nieistotnych znaków lub zamianę ich na równoważne (Delgado, Araki, Neto, 2005, s. 16–30).

Następnym krokiem jest analiza morfologiczna, tj. podział tekstu na zdania, wyrazy oraz sylaby. Ostatnim krokiem normalizacji jest translacja przetworzonych danych tekstowych na alfabet fonetyczny X-SAMPA lub IPA (zaimplementowano w badaniach). Poprawność działania wyżej opisanych algorytmów ma kluczowe znaczenie dla syntezy mowy, gdyż wyniki ich pracy będą prze-

kazywane bezpośrednio do metody odpowiedzialnej za syntezę mowy, która będzie „czytać" tekst użytkownikowi (Graliński, Jassem, Wagner, Wypych, 2006, s. 10).

Szybkość działania mechanizmów normalizacji treści

Szybkość działania metody normalizacji tekstu ma kluczowy wpływ na działanie całej syntezy, gdyż zbyt długie przetwarzanie tekstu wydłuży czas oczekiwania na dźwięk, co w konsekwencji może prowadzić do tego, że urządzenie będzie czytało napisy, które już nie są istotne dla użytkownika.

Programy testowe zaimplementowano w języku C# wraz z graficznym interfejsem użytkownika. Do sprawdzenia wydajności przygotowano 10 tekstów różnej długości, zawierających elementy wymagające translacji lub usunięcia (liczby arabskie, rzymskie, daty, godziny, skróty oraz znaki interpunkcyjne).

Uzyskano średni czas normalizacji ok. 15 ms, przy czym należy zauważyć, że w badaniu wykorzystano kilka długich fragmentów tekstu (1700 znaków – 3 akapity tekstu).

Podczas analizy wydajności algorytmu ważnym elementem jest również koszt obliczeniowy danej metody. W projekcie ToucanEye ma to szczególne znaczenie, gdyż urządzenie będzie kompaktowych rozmiarów, a zasilanie będzie bateryjne. Średnie zużycie procesora dla ww. przykładów to 12,5%, a średnie wykorzystanie pamięci RAM to 60,5 MB.

Podsumowując badania wydajności mechanizmów i algorytmów normalizacji tekstu, można stwierdzić, że czas potrzebny na przetwarzanie treści oraz wykorzystanie zasobów jest krótki i nie powinien mieć negatywnego wpływu na działanie urządzenia.

Poprawność normalizacji treści

Przygotowanie tekstu, który następnie zostanie przekazany do syntezy mowy, nie jest zagadnieniem trywialnym i od jakości tego przygotowania w dużej mierze zależy jakość i zrozumiałość mowy. Elementy podlegające przekształceniom to liczby naturalne, liczby rzeczywiste, liczby poprzedzone lub zakończone symbolami, daty, godziny, symbole i znaki, skróty i skrótowce, adresy e-mail, strony WWW, znaki tabulacji lub końca linii.

Tabela 1. Poprawność działania poszczególnych metod translacji

Daty	97,5%
Godziny	94,8%
Liczby całkowite	100%
Liczby zmiennoprzecinkowe	100%
Liczby rzymskie	99,9%

Źródło: opracowanie własne.

Do przeprowadzenia badania wygenerowano bazę 500 fraz testowych dla każdej metody translacyjnej. W przypadku dat frazy posiadały różne sposoby zapisu daty (np. data w formacie dd-mm-yyyy, dd/mm/yyyy, dd.mm.yyyy i różnego rodzaju kombinacje). Wyniki badania skuteczności pokazano w tabeli 1.

Powyższe badanie miało na celu sprawdzenie poprawności działania poszczególnych metod, lecz problem normalizacji treści jest o wiele bardziej złożony niż prosta zamiana znaków niebędących literami na ich słowną interpretację. Zadaniem nietrywialnym w normalizacji tekstu dla języka polskiego jest poprawne wykrycie formy dla danego wyrazu. Na przykład w zdaniu „Ala ma 17 kotów" wymagana jest zmiana zapisu liczby 17 na „siedemnaście", które w tym wypadku jest w formie podstawowej. Natomiast w zdaniu „Ukończył bieg na 10 pozycji" liczba 10 powinna zostać zamieniona na „dziesiątej". Kolejnym przykładem jest „Widziałem ich 14 lipca wieczorem w Gdańsku", gdzie liczba 14 powinna zostać znormalizowana do zapisu czternastego (Perkins, 2014).

Powyższe trzy przykłady pokazują, że zamiana liczby na formę pisaną w odpowiedniej formie dla kontekstu zdania nie jest prostym zagadnieniem. Skróty podobnie jak liczby wymagają zapisu w innej formie niż podstawowa. Zdanie „Na zewnątrz był tylko 1 °C", gdzie '°C' powinno być zamienione na 'stopień Celsjusza', gdzie przeważnie stosowana jest forma stopni Celsjusza. Kolejnym przykładem jest zdanie „Muzeum jest czynne od godz. 8:00", gdzie wymagana jest zamiana skrótu 'godz.' na 'godziny' oraz ciągu znaków '8:00' na 'ósmej'. Ostatnim przykładowym zdaniem jest „Samochód ten został wyprodukowany 15/04/2004 roku" gdzie 15/04/2004 powinno zostać zamienione na 'piętnastego kwietnia dwa tysiące czwartego'. Przykładowe zdania służące do testu pokazano w tabeli 2.

Tabela 2. Przebieg badania poprawności translacji tekstu

Zdanie testowe	Zdanie po translacji	Wynik
W budynku znajdują się 82 mieszkania oznaczone numerami od 1 do 82.	W budynku znajdują się osiemdziesiąt dwa mieszkania oznaczone numerami od jeden do osiemdziesiąt dwa.	1
Woda nie ma stałej temperatury wrzenia, a wrzenie w 100 °C występuje tylko w tzw. warunkach normalnych.	Woda nie ma stałej temperatury wrzenia, a wrzenie w sto stopień Celsjusza występuje tylko w tak zwany warunkach normalnych.	0

Źródło: opracowanie własne.

Testy przeprowadzono na zbiorze 50 zdań zawierających elementy wymagające tłumaczenia (liczby, daty, godziny, skróty). Elementy w zdaniach przetłumaczone bezbłędnie otrzymały ocenę 1, natomiast zdanie, w którym element został źle przetłumaczony lub forma tłumaczenia była zła, otrzymało ocenę 0. Wynik badania poprawności translacji to 55%.

Synteza mowy

Badania zostały przeprowadzone dla systemu MBROLA oraz Festival Speech Synthesis Systems. Dla uzyskania możliwie obiektywnych wyników w obu przypadkach użyto bazy mowy Krzysztofa Szklannego (dostępną zarówno dla systemu MBROLA oraz Festival). Pomiary szybkości generowania mowy podzielono na kilka kategorii w zależności od długości syntezowanego tekstu. Przykładowe dane testowe:

– 1–3 słowa: „zastosowaniem programu mowy",

– 3–7 słów: „Dane wejściowe systemu klasyfikacji stanowią wektory",

– 7–10 słów: „łączenie wypowiedzi z nagranych fragmentów głosu lektora zawierających słowa",

– 10–20 słów: „Ostatni etap przewiduje projekt i budowę funkcjonalnej obudowy dla urządzenia. Wybranie optymalnej formy urządzenia może wymagać stworzenia kilku projektów".

Metoda konkatenacyjna

Synteza mowy konkatenacyjnej generuje mowę poprzez sklejanie ze sobą elementów akustycznych powstałych z naturalnej mowy (fony, difony, trifony, sylaby). Badanie metody konkatenacyjnej polegało na wyodrębnieniu z korpusu mowy bazy zawierającej pojedyncze wystąpienie każdej jednostki akustycznej (NKJP, 2017).

Czasy uzyskane w tej metodzie pokazano w tabeli 3.

Tabela 3. Czas oczekiwania na syntezę [sek.]

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,052	0,049	0,096
3–7 słów	0,095	0,094	0,143
7–10 słów	0,181	0,176	0,257
10–20 słów	0,322	0,297	0,448

Źródło: opracowanie własne.

Kolejnym krokiem była modyfikacja algorytmów syntezy poprzez multiplikację wątków zajmujących się procesem syntezy. Multiplikacja wątków odpowiedzialnych za generowanie mowy nie dała pozytywnych rezultatów i w przeciwieństwie do metody korpusowej spowodowała znaczne wydłużenie czasów. Czasy uzyskane za pomocą tej modyfikacji pokazano w tabeli 4.

Dużą zaletą metody konkatenacyjnej jest niewielki rozmiar bazy danych, co implikuje mniejsze zapotrzebowanie na zasoby sprzętowe niż w przypadku innych metod. Średnia zasobochłonność procesu syntezy to 4,47% użycia procesora oraz 46 MB użycia pamięci RAM.

Tabela 4. Czas oczekiwania [sek.] w na syntezę przy rozdzieleniu procesu syntezy mowy na kilka procesów

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,167	0,276	0,666
3–7 słów	0,158	0,386	1,061
7–10 słów	0,232	0,550	1,764
10–20 słów	0,392	0,788	2,762

Źródło: opracowanie własne.

Metoda korpusowa

W metodzie konkatenacyjnej występuje jedna postać jednostki akustycznej, natomiast metoda korpusowa to modyfikacja metody konkatenacyjnej, gdzie występuje wiele postaci tej samej jednostki akustycznej. Korpus jest dużo większy, tak że zawiera po kilka instancji danego difonu. W celu wygenerowania mowy obliczana jest funkcja kosztu, która polega na obliczeniu optymalnego połączenia jednostek mowy dla uzyskania możliwie najlepszej jakości mowy (NKJP, 2017).

MBROLA

Jest to wielojęzyczny system syntezy mowy stworzony na Polytechnique de Mons w Belgii. Badania systemu zostały przeprowadzone dla korpusu Krzysztofa Szklannego. Tabela 5 przedstawia wyniki uzyskane dla systemu MBROLA (MBROLA, 2017).

Tabela 5. Czas oczekiwania [sek.] na syntezę w systemie MBROLA [sek.]

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,015	0,020	0,066
3–7 słów	0,016	0,019	0,089
7–10 słów	0,017	0,021	0,117
10–20 słów	0,017	0,039	0,1571

Źródło: opracowanie własne.

Średnia zasobochłonność procesu syntezy przez MBROLĘ to 5% użycia procesora oraz 22 MB użycia pamięci RAM.

FESTIVAL

Jest to wielojęzyczny system syntezy mowy stworzony na Uniwersytecie w Edynburgu w Centrum Rozwoju Technologii Mowy. Badania systemu zostały przeprowadzone dla dwóch polskich korpusów mowy (Festival, 2017).

Tabela 6. Czas oczekiwania [sek.] na syntezę z wykorzystaniem bazy cstr_pl

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,023	0,058	0,043
3–7 słów	0,033	0,052	0,068
7–10 słów	0,063	0,095	0,119
10–20 słów	0,103	0,177	0,213

Źródło: opracowanie własne.

Słownik Krzysztofa Szklannego

Podczas badań zauważono, że czas na odpowiedź systemu wzrosła znacząco w porównaniu do korpusu ‘cstr_pl’ i może trwać kilka, a nawet kilkanaście sekund. Jakość mowy generowana z użyciem dużego korpusu mowy znacząco przewyższała tę związaną z podstawowym słownikiem, w związku z czym podjęto prace nad optymalizacją algorytmów (PJWSTK, 2017).

Przeprowadzono optymalizację szybkości działania mechanizmu poprzez zwiększenie liczby procesów odpowiedzialnych za działanie systemu FESTIVAL. Działanie takie pozwoliło na znaczne skrócenie szybkości syntezy, lecz czas ten nadal przekraczał 2 sekundy. W związku z tym podjęto badania nad doborem szybszych metod doboru jednostki i kosztu obliczeniowego konkatenacji. W późniejszych krokach ograniczono bazę korpusu mowy, co pozwoliło na znaczne przyspieszenie czasu generowania mowy. W tabelach 7–11 przedstawiono czasy generowania mowy dla kolejno modyfikowanych baz.

Tabela 7. Czas oczekiwania [sek.] na syntezę dla pełnego korpusu mowy (2050 zdań)

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,521	0,793	1,410
3–7 słów	0,855	1,380	2,354
7–10 słów	1,569	2,410	4,201
10–20 słów	3,020	3,798	7,445

Źródło: opracowanie własne.

Tabela 8. Czas oczekiwania [sek.] na syntezę dla okrojonego korpusu mowy (1050 zdań)

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,313	0,351	0,576
3–7 słów	0,549	0,596	1,144
7–10 słów	1,006	1,144	2,09
10–20 słów	1,723	1,64	3,729

Źródło: opracowanie własne.

Tabela 9. Czas oczekiwania [sek.] na syntezę dla okrojonego korpusu mowy (525 zdań)

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,171	0,230	0,335
3–7 słów	0,297	0,454	0,672
7–10 słów	0,550	0,911	1,367
10–20 słów	0,894	0,842	2,083

Źródło: opracowanie własne.

Tabela 10. Czas oczekiwania [sek.] na syntezę dla co najmniej 4 wystąpień każdego z fonemów

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,216	0,370	0,372
3–7 słów	0,393	0,711	0,739
7–10 słów	0,741	0,959	1,460
10–20 słów	1,228	1,618	2,568

Źródło: opracowanie własne.

Tabela 11. Czas oczekiwania [sek.] na syntezę dla co najmniej 7 wystąpień każdego z fonemów

	Urządzenie 1	Urządzenie 2	Urządzenie 3
1–3 słowa	0,288	0,439	0,540
3–7 słów	0,479	0,654	0,983
7–10 słów	0,965	1,192	1,986
10–20 słów	1,663	2,094	3,573

Źródło: opracowanie własne.

Średnia zasobochłonność systemu FESTIVAL to 14,81% użycia procesora oraz 276,95 MB użycia pamięci RAM.

Zrozumiałość syntezy mowy

W celu sprawdzenia jakości syntezowanej mowy przeprowadzono testy odsłuchowe. Eksperci mieli za zadanie ocenić zrozumiałość i naturalność generowanych przez poszczególne metody i mechanizmy syntezy wypowiedzi.

Do oceny jakości syntezy mowy został wykorzystany współczynnik MOS (ang. *Mean Opinion Score*) używany w telefonii do oceny jakości dźwięku po kompresji, dekompresji lub transmisji, gdzie zasady określenia MOS zostały unormowane przez ITU (Międzynarodowy Związek Telekomunikacyjny) w zalecaniu ITU-T P.800.s Do określenia wartości posłużyła ocena ACR (ang. *Absolute Category Rating*).

Punktem bazowym dla wyników pracy jest średnia ocena MOS dla systemu GPS FR (260 bitów * 50 próbek/s – 13 kbit/s, algorytm kodujący / dekodujący Regular Pulse Excitation – Long Term Prediction Linear Predictive Coder [RPE-LTP]) wynoszącej 3,5.

Tabela 12. Współczynnik MOS dla różnych metod syntezy

MBROLA	espeak-PL	espeak PL + F2	festival metoda konkatenacyjna	festival baza zawie-rająca 4 wystąpienia difonów	festival baza zawie-rająca 7 wystąpień difonów	festival baza 525 fraz	festival baza 1050 fraz	festival baza 2050 fraz	festival baza cstr_pl
1	2	3	4	4	4	4	4	5	2
2	3	3	4	5	4	4	4	5	2
2	2	3	4	4	5	5	5	5	2
2	2	3	4	4	4	4	4	5	3
2	3	3	4	4	5	4	5	5	2
3	2	3	4	4	4	4	4	5	2
3	2	3	4	4	4	4	5	5	2
2	2	3	5	5	5	5	5	5	2
1	1	1	3	4	4	4	5	5	3
2	3	3	4	4	5	4	5	5	3
2	3	3	3	4	4	4	4	4	2
Średnia:									
2,08	**2,25**	**2,75**	**3,92**	**4,17**	**4,33**	**4,17**	**4,50**	**4,83**	**2,25**

Źródło: opracowanie własne.

Dla porównania jakości generowanej mowy do testu dodano również pliki wygenerowane przez program ‘eSpeak text to speech’ z dwoma polskimi słownikami. Wyniki dla 12 badanych osób pokazano w tabeli 12.

Skala MOS przyjmuje wartości od 1 do 5:

– 1 – zła jakość, zrozumienie mowy jest niemożliwe,
– 2 – słaba jakość, zrozumienie mowy jest bardzo trudne, uciążliwe dla słuchacza,
– 3 – średnia jakość, zrozumienie mowy jest możliwe, lecz wymaga koncentracji,
– 4 – dobra jakość, występują błędy, lecz nie wpływają one na zrozumienie,
– 5 – znakomita jakość, pełne zrozumienie mowy.

Podsumowanie

Przeanalizowano metody przetwarzania tekstu w języku polskim. Szczególnie skupiono się na płytkiej analizie tekstu. Szerzej zbadano algorytmy translacji znaków niebędących literami i skrótów na postać słowną oraz metody translacji tekstów na alfabety fonetyczne X-SAMPA oraz IPA. Przeprowadzono testy, podczas których wykazano, że poprawność translacji kształtuje się na poziomie 55%.

Przeanalizowano dostępne metody syntezy mowy pod względem szybkości generowania, a także zrozumiałości i naturalności mowy. Opisywana metoda konkatenacyjna pozwala na generowanie dobrej jakości mowy. Potwierdzają to badania MOS dla wszystkich metod, gdzie otrzymano średnią ocenę 3,92. Czas generowania mowy na poziomie pierwszej wypowiedzi wynosi poniżej 0,5 sekundy, co czyni tę metodę optymalnym kandydatem na wykorzystanie w urządzeniu przenośnym. Dalsze prace nad opisanymi metodami pozwolą na ich optymalizację i poprawę jakości działania.

Literatura

Delgado, R., Araki, M., Neto, J. (2005). *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assesment*. Richmond: Wiley.

FESTIVAL. Pobrane z: http://www.cstr.ed.ac.uk/projects/festival/ (1.09.2017).

Graliński, J., Jassem, K., Wagner, A., Wypych, M. (2006). Linguistic Aspects of Text Normalization in Polish Text-to-speech System. *System Science*, *32* (4), 7–15.

Łopatka, K., Czyżewski, A. (2010). Syntetyzer mowy uwzględniający prozodię wypowiedzi. *Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej*, *28*, 105–108.

MBROLA (2017). Pobrane z: http://tcts.fpms.ac.be/synthesis/mbrola.html (1.09.2017).

NKJP (2017). Pobrane z: http://nkjp.pl/ (1.09.2017).

Perkins, J. (2014). *Python 3 Text Processing with NLTK 3 Cookbook*. Birmingham: Packt Publishing.

PJWSTK (2017). Pobrane z: http://syntezamowy.pjwstk.edu.pl/synteza.html (1.09.2017).

Tadeusiewicz, R. (1988). *Sygnał mowy*. Warszawa: Wydawnictwa Komunikacji i Łączności.