AKADEMIA GÓRNICZO-HUTNICZA

IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Informatyki, Elektroniki i Telekomunikacji

Katedra Elektroniki

PRACA DYPLOMOWA

Inżynierska

Temat: ***Korpus mowy telefonicznej***

Telephone speech corpus

Imię i nazwisko: Robert Dyjas

Kierunek studiów: Elektronika i Telekomunikacja

Opiekun pracy: dr inż. Bartosz Ziółko

Oświadczam, świadomy odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem osobiście i samodzielnie i że nie korzystałem ze źródeł innych niż wymienione w pracy.

Spis treści

Wstęp do pracy

Poniższy rozdział ma za zadanie przedstawienie celu oraz zakresu pracy dyplomowej. Zostały w nim omówione wymagania stawiane dyplomantowi oraz przedstawione etapy, z których składa się praca.

Cel pracy

Celem pracy jest stworzenie korpusu mowy polskiej. Ma on spełniać szereg wymagań takich jak:

- Minimalna sumaryczna długość nagrań: 100 minut
- Treść nagrań ma zawierać przede wszystkim losowe ciągi cyfr/liczb w różnych konfiguracjach
- Jeżeli ta sama osoba czytająca ma odczytać więcej niż jeden zestaw ciągów, to zestawy te mają być dopasowane w sposób zapewniający jak największą różnorodność
- Sugerowane jest, by korpus zawierał krótsze wypowiedzi większej liczby osób niż długie wypowiedzi mniejszej liczby osób
- Dla zapewnienia realności nagrań mają one być zapisem rozmowy telefonicznej żeby uwzględnić wszelkie rodzaje zniekształceń dźwięku powstałe na linii nadawca – odbiornik
- Do każdego nagrania mają być zebrane odpowiednie metadane opisane szczegółowo w podrozdziale 2.4
- Zebrane dane powinny być „autentyczne”, tj. np. jeśli zbierane są dane o strukturze numeru PESEL, to powinny one odpowiadać istniejącym aktualnie numerom PESEL

Zakres pracy

Zakres pracy obejmuje cały proces powstawania korpusu, począwszy od przygotowania danych, które potem mają stanowić treść nagrań. Powinny one spełniać wspomniany wyżej warunek „autentyczności”, więc w tym celu należy dokonać sprawdzenia. Następnym krokiem jest odpowiednie dopasowanie danych. Należy więc znaleźć lub stworzyć algorytm, dzięki któremu zostanie osiągnięta maksymalna możliwa różnorodność nagrań.

Stworzenie systemu umożliwiającego automatyczne lub półautomatyczne przeprowadzenie samego procesu nagrywania jest ostatnim z kroków poprzedzających samo nagrywanie wypowiedzi. Zadaniem dyplomanta jest znalezienie odpowiedniej liczby osób chętnych nagrać swoje wypowiedzi, odpowiednie rozdystrybuowanie do tych osób treści oraz objaśnienie im procesu nagrywania. Ze względu na zachowanie naturalności wskazane jest również, by w miarę możliwości nie sugerować osobom nagrywającym wypowiedzi sposobu odczytywania ciągów (zarówno jeśli mowa o sposobie odczytywania dat, jak i o sposobie składania ze sobą cyfr).

Ostatnim etapem procesu przygotowania korpusu mowy jest odpowiednia obróbka tych danych. W pojęciu tym mieści się najpierw zebranie wszystkich nagrań oraz ich poprawna anotacja. Obróbka obejmuje również przygotowanie słowników do wszystkich nagrań oraz dostarczenie ich do Zespołu Przetwarzania Sygnałów celem transkrypcji przy użyciu systemu „AGH SARMATA" (więcej o systemie w rozdziale 1.7). Dane wyjściowe z tego systemu mają zostać również przekonwertowane przez dyplomanta do formatu .mlf, który został szczegółowo opisany w rozdziale 1.8.

1. Opracowanie teoretyczne

1.1. Podstawowe pojęcia

Korpus mowy - zbiór tekstów będący podstawą badań językoznawczych [1]. W opracowaniu, jako korpus, traktowany jest zbiór: tekstów, nagrań tychże tekstów na nośnik audio (plik .wav), anotacji (zawierającej również metadane) oraz pliki do tych nagrań w formacie .mlf.

Anotacja (ang. annotation) stanowi dodatkową warstwę informacyjną, dodaną przez twórców korpusu, umożliwiającą tworzenie bardziej precyzyjnych zapytań oraz wyszukiwanie informacji dodatkowych nie zawartych bezpośrednio w zbiorze tekstów źródłowych. Anotacja może również obejmować metadane i pozwalać na wyszukiwanie, np. według daty powstania lub płci autora tekstu. Zarówno w tekstach anglo- jak i polskojęzycznych, termin anotacja jest używany zamiennie z terminem znakowanie (ang. mark-up), choć niektórzy autorzy postulują rozróżnienie tych dwóch terminów [2]. W opracowaniu tym przyjęto terminy anotacja oraz znakowanie jako jednoznaczne.

Znakowanie – j.w.

Plik .mlf – plik zawierający informacje odnośnie początku i końca trwania frazy/wyrazu/głoski. W niniejszym opracowaniu pliki te będą zawierać informacje odnośnie każdego wyrazu.

Rozpoznawanie mowy (ang. ASR - automatic speech recognition) – technologia pozwalająca komputerowi lub innemu urządzeniu interpretować mowę ludzką, na przykład do celów transkrypcji lub jako alternatywną metodę interakcji.

Fonem - [gr. phṓnēma ‘dźwięk’], podstawowa jednostka fonologiczna stanowiąca teoretyczną abstrakcję w stosunku do głosek [3].

1.2. Historia rozpoznawania mowy [4] [5]

Zainteresowanie badaniami nad ludzką mową nie jest zjawiskiem nowym. Słowa wypowiadane były przez długi czas głównym medium w komunikacji międzyludzkiej. Stąd ciekawość ludzi nauki skupiała się zarówno na mechanicznej syntezie mowy, jak i na rozpoznawaniu wypowiedzianych słów celem zautomatyzowania niektórych czynności i prostych zadań.

Samo zainteresowanie wyszukiwaniem sposobów na bardziej efektywne wykonywanie różnych czynności było jednym z powodów, dla których powstały pierwsze gramofony. Miały one na celu stworzenie możliwości do nagrania na nośnik

wiadomości lub listu, który potem miał być przekazany sekretarce celem przepisania. Dzięki zapisaniu takich danych na nośniku, można było oszczędzić na zatrudnieniu kosztownych stenografów.

Samo rozpoznawanie mowy było tematem poruszanym w kilku filmach z okresu lat sześćdziesiątych i siedemdziesiątych. Przykładem może być film „2001: Odyseja kosmiczna". Został tam uruchomiony robot HAL, który potrafił rozpoznawać kierowane do niego zdania oraz na nie odpowiadać.

Pierwsze próby rozpoznawania mowy opierały się na rozpoznawaniu pojedynczych dźwięków na podstawie widma mocy sygnału głosowego. W 1952 roku Davis, Biddulph oraz Balashek z Bell Laboratories zbudowali system pozwalający na rozpoznanie pojedynczych cyfr. Do jego poprawnego działania wymagana była jednak wcześniejsza detekcja parametrów głosu konkretnego mówcy oraz założenie o niezmienieniu swojej pozycji względem mikrofonu podczas testowania i detekcji.

W latach sześćdziesiątych poprzedniego wieku została opracowana szybka transformata Fouriera (ang. FFT – Fast Fourier Transform) oraz niejawne łańcuchy Markowa (ang. HMM – Hidden Markov Model). Oba te narzędzia są wykorzystywane również w obecnych badaniach nad przetwarzaniem mowy. Wspomniane już lata sześćdziesiąte przyniosły również, uznawany za pierwszy, pełny syntezator mowy wykonany przez Noriko Umeda oraz jego zespół w 1968 roku. Syntezator ten działał tylko dla języka angielskiego.

Na skutek pięcioletniego projektu finansowanego przez ARPA „Speech Understanding Project" powstał we wrześniu 1976 roku system „CMU Harpy" [6]. Umożliwiał on rozpoznawanie 1000 słów ze słownika z dokładnością powyżej 90%.

1.3. Rozpoznawanie mowy polskiej

Wg artykułu George'a H.J. Weber'a [7] w 1997 roku językiem angielskim posługiwało się aż 480 mln ludzi, zaś język polski 10 lat później był głównym językiem dla zaledwie 40 milionów ludzi [8]. Podane liczby uzasadniają mniejszy nacisk badaczy skierowany na pracę nad rozpoznawaniem mowy polskiej.

Jednakże, w ostatnich latach, pojawiły się wdrożenia systemów rozpoznawania mowy na dużą skalę. Należy do nich między innym trwający od 2013 roku projekt prowadzony w ramach Polskiej Platformy Bezpieczeństwa Wewnętrznego [9], program SkryBot, Google Web Speech API (w wersji beta) [10] oraz system rozpoznawania mowy

polskiej SARMATA, stworzony przez Zespół Przetwarzania Sygnałów Akademii Górniczo-Hutniczej (opisany szerzej w rozdziale 1.7).

1.4. Zastosowania korpusów

Wszelkie systemy rozpoznawania mowy potrzebują do swojego działania znać dźwięki/głoski/wyrazy, które mają rozpoznawać. W tym celu potrzebują tzw. danych treningowych. Danych tego typu może dostarczyć np. korpus mowy.

Odpowiednie i precyzyjne wykonanie korpusu jest bardzo ważnym elementem wykonania sprawnie działającego systemu rozpoznawania mowy. Dzięki starannie przygotowanym danym treningowym, system może w bardziej efektywny sposób nauczyć się podanych mu dźwięków, co przekłada się na jego większą dokładność.

W zależności od docelowego zadania, które zostanie postawione systemowi ASR, korpus, który będzie go trenował, może (a czasami wręcz powinien) się zdecydowanie różnić od innych korpusów występujących dla danego języka. Przykładowo, przy tworzeniu systemu dla wymiaru sprawiedliwości w słowniku powinno występować słowo „repertorium", które w interfejsie obsługi, np. komputera, nie jest wcale używane (przykład takiego interfejsu można zobaczyć w [11]).

Korpus mowy nie musi się wcale ograniczać do zwykłej mowy. Pokazują to przykłady korpusów opracowanych w Zespole Przetwarzania Sygnałów AGH (ich opisy są dostępne na stronie internetowej [12]). Grupa ta opracowała m.in. korpus audiowizualny oraz korpus emocji w mowie.

W niniejszym opracowaniu został przygotowany korpus, którego dane w dużej mierze opierają się na cyfrach i liczbach, jak również datach. Może on być wykorzystany na przykład przy tworzeniu systemu, który pozwoli na automatyczne sterowanie jakimś systemem, którego dane wejściowe to liczby. Dane tekstowe, które zostały przygotowane do przeczytania, mają strukturę ciągów, którymi posługujemy się coraz częściej w rozmowach telefonicznych, jak i w codziennym życiu. W dzisiejszych czasach każdy bank czy operator telefonii komórkowej oraz niektóre z firm, świadczących usługi na podstawie umów zawieranych drogą telefoniczną, posiadają systemy weryfikacji rozmówców. Niestety, (zarówno dla klientów jak i świadczących usługi) często weryfikacja ta polega na żmudnym powtarzaniu wszystkich lub części swoich danych osobowych. System rozpoznawania mowy oparty na przygotowanym przez dyplomanta korpusie, może zdecydowanie usprawnić ten element.

1.5. Wybrane korpusy mowy polskiej

Zdecydowanie najpopularniejszym z korpusów mowy polskiej jest CORPORA autorstwa Stefana Grocholewskiego. Została ona wykonana w 1997 roku na Politechnice Poznańskiej. Sam autor opisuje ten korpus w następujący sposób: "Dla każdego z 45 mówców dokonano nagrań 365 wypowiedzi. Do nagrań wykorzystano mikrofony pojemnościowe lub w jednym przypadku mikrofon dynamiczny. Parametry nagrań: częstotliwość próbkowania - 16 kHz, długość słów - 12 bitów. Nagrań dokonano w warunkach naturalnych pomieszczeń, w bezpośredniej bliskości pracującego komputera" [13].

Sama treść korpusu autorstwa p. Grocholewskiego na pierwszy rzut oka może się wydawać bezsensowna, ponieważ została ona dobrana pod kątem zapewnienia jak największej różnorodności fonetycznej. Stąd zdania takie, jak: „lubić czardaszowy pląs" czy „on myje wróble w zoo". Niemal wszystkie wypowiedzi, (oprócz 2 mówców) zostały zanotowane automatycznie właśnie na podstawie ręcznie posegmentowanych wypowiedzi wspomnianej dwójki mówców. Był to jeden mężczyzna oraz jedna kobieta [14].

Kolejny ważny korpus mowy polskiej to jurisdic. Zawiera on nagrania o tematyce prawniczej. Nagrane są zarówno spontaniczne wypowiedzi, jak i teksty czytane. Wedle relacji autorów zawiera on około 1000 mówców z różnych części Polski [15].

Korpus LUNA zawiera dialogi telefoniczne. Został on stworzony celem opracowania narzędzia do usprawnienia obsługi serwisów telefonicznych. Korpus zawiera zarówno rozmowy człowieka z człowiekiem, jak i człowieka z komputerem.

Narodowy Korpus Języka Polskiego jest zbiorem półtora miliarda słów, zaczerpniętych z literatury, mediów, listów, tekstów internetowych itp. Powstawał on w latach 2008-2012 przy współpracy Polskiej Akademii Nauk, Wydawnictwa Naukowego PWN oraz Uniwersytetu Łódzkiego. Był projektem badawczym Ministerstwa Nauki i Szkolnictwa Wyższego. Posiada on swoją stronę, na której można zarówno przeczytać dodatkowe informacje o nim jak i przeszukać jego zawartość. Był on również wykorzystywany w projektach przeprowadzonych na AGH [16].

Część korpusu stanowią nagrania rozmów i audycji radiowych, które niestety nie są zanotowane czasowo [17].

1.6. Korpusy powstałe na AGH [17]

Najważniejszym korpusem powstałym na AGH jest Korpus AGH zawierający ponad 25 godzin nagrań. Są to nagrania 166 mówców, głównie w przedziale wiekowym 20-35 lat. Większość nagrywających to mężczyźni.

Nagrania w korpusie to dźwięki w formacie .wav jednokanałowe. Część z nagrań została zanotowana ręcznie, a część za pomocą OpenSJP (dystrybuowanego na licencji open source słownika języka polskiego) oraz ręcznie poprawiona. Niektóre słowa w tym korpusie (np. zapożyczone z innych języków) zostały przekonwertowane za pomocą oprogramowania ORTFON [18]. Korpus zawiera też zasady, za pomocą których słowa te zostały przetworzone, co umożliwia zastosowanie tego procesu w drugą stronę.

10 godzin z nagrań stanowią nagrania języka potocznego. Zostały one wykonane przez 10 osób. Każda z nich czytała około 1000 zwrotów. Nagrania odbywały się w cichym pokoju. Anotacja została wykonana na poziomie całych zwrotów.

Niemal 7 godzin stanowią nagrania wykonane przez studentów podczas zajęć z przedmiotu prowadzonego przez Zespół Przetwarzania Sygnałów. Warunkiem zaliczenia przedmiotu było wykonanie prostego systemu rozpoznawania mowy o dowolnej tematyce. Najczęściej był to system obsługujący zamówienie pizzy, kupno biletu autobusowego lub stworzenie interfejsu do obsługi aplikacji. Każdy system składał się z około 3 minut nagrań. Dotychczas takie zadanie wykonało 125 studentów, głównie w wieku 20-25 lat. Proporcje odnośnie płci to dwóch mężczyzn na jedną kobietę.

Jedna z części korpusu została przygotowana specjalnie pod kątem przygotowanie systemu służącego do syntezy tekstu na mowę. Część ta składa się z 2132 zdań przeczytanych przez młodą kobietę. Tekst został przygotowany na podstawie NKJP (Narodowego Korpusu Języka Polskiego) [19] i dobrany pod kątem zapewnienia jak największej różnorodności fonetycznej, i jak największego podobieństwa do języka mówionego. Całe 4 i pół godziny nagrań zostało przygotowanych za pomocą wysokiej jakości sprzętu z wykorzystaniem komory bezechowej.

Subkorpus przygotowany na podstawie nagrań VOIP został wykonany przez 27 mówców, głównie w wieku 20-35 lat. Zawiera on niemalże 3 godziny nagrań, których zawartość to w głównej mierze cyfry i liczby oraz słowa służące do nawigacji głosowej telefonicznego systemu pomocy technicznej. Anotacja tego subkorpusu została wykonana na poziomie słów.

Pozostałe nagrania wykonane na potrzeby korpusu stanowią m.in. nagrane komendy służące do sterowania systemem ASR SARMATA (więcej o systemie w rozdziale 1.7) oraz używane w systemie SAWA (interfejs głosowy wykonany dla instytucji wymiaru sprawiedliwości). Nagrane są publiczne wykłady oraz prezentacje wykonane przez członków Zespołu Przetwarzania Sygnałów AGH. Nagrania te trwają w sumie nieco ponad 1,5 godziny. Są zanotowane na poziomie słów.

Kolejny ciekawy korpus wykonany przez zespół DSP AGH to audiowizualny korpus mowy. Zawiera on ponad 3 godziny nagrań głosu oraz twarzy (patrząc od przodu). Nagrania zostały zarejestrowane w rozdzielczości Full HD, głównie przy naturalnym oświetleniu. Przedstawiony korpus może służyć na przykład do trenowania systemu służącego do rozpoznawania mowy z ruchu warg [20].

Korpus emocji w mowie jest jedynym korpusem w Polsce, który zawiera dostępne w ramach licencji nagrania ludzkich emocji. Zawiera on 6 różnych emocji oraz stan neutralny, jako sygnał odniesienia. Swojego głosu użyczyło 12 mówców; zarówno profesjonalnych aktorów, jak i studentów. Każde nagranie zawiera tę samą treść, a mówca został poproszony o przeczytanie konkretnych fraz w sposób wyrażający daną emocję. Korpus zawiera zarówno tekst ciągły, jak i cyfry, polecenia oraz zdania [21].

1.7. SARMATA

System rozpoznawania mowy polskiej SARMATA jest kolejnym z projektów wykonanych przez wspomniany już Zespół Przetwarzania Sygnałów (DSP) AGH. Jest aplikacją przystosowaną do obsługi do 1000 komend jako np. interfejs głosowy.

Zastosowania SARMATY to, między innymi, prowadzenie interaktywnych rozmów (ang. IVR – Interactive Voice Response), za pomocą których można zautomatyzować działanie systemu call center. W połączeniu z innym systemem weryfikacji mówców (jak np. Voice Color spółki Techmo, będącej spin-offem Akademii Górniczo-Hutniczej w Krakowie [22]) może on odciążyć pracownika od odpowiadania na powtarzające się pytania i tym samym zmniejszyć koszt biznesowy utrzymania call center.

Jak zostało już wcześniej wspomniane, SARMATA może służyć jako system to głosowej obsługi programów komputerowych, co oprócz korzyści polegających na zwiększeniu efektywności pracy, może być też znaczącym ułatwieniem dla niepełnosprawnych, np. słabowidzących i mających trudności w posługiwaniu się myszą.

Przykładem takiego właśnie interfejsu jest wykonana za pomocą SARMATY Wirtualna Mysz. Jest to system, który pozwala na sterowanie myszą komputerową za pomocą głosu.

Kolejnym z zastosowań systemu SARMATA jest wyszukiwarka akustyczna. Za jej pomocą możliwe jest odszukanie w nagraniach danego słowa. Pozwala to znacznie zaoszczędzić czas wszystkim osobom, które pracują analizując pliki dźwiękowe i wyszukując w nich odpowiednie dane.

W niniejszym opracowaniu opisany powyżej system został użyty do analizy nagrań będących składową korpusu. Dzięki jego wykorzystaniu został zaoszczędzony czas, który musiałby zostać poświęcony na ręczną transkrypcję dźwięków. Przy założeniu, że jedna minuta nagrań wymaga poświęcenia około 20 minut na ręczną transkrypcję, dzięki SARMACIE zaoszczędzonych zostało ponad 40 godzin pracy.

1.8. Format .mlf

Pliki w formacie .mlf (Master Label File) służą do zapisania danych powstałych w procesie transkrypcji. W zależności od dokładności transkrypcji zawierają one informacje o fonemie/słowie/frazie oraz ich czasie, w którym dana jednostka się zaczyna i kończy.

Plik .mlf musi zaczynać się nagłówkiem #!MLF!#. W zależności od potrzeby może on zawierać anotację dowolnej liczby plików. Każda z nich musi być poprzedzona ścieżką do pliku dźwiękowego wziętą w cudzysłów. Po podaniu ścieżki następują dane zapisane w formacie: czas rozpoczęcia, czas zakończenia, treść (rozdzielonej spacjami). Podstawową jednostką czasu w pliku .mlf jest 100 ns, więc w przypadku uzyskania anotacji z większą jednostką podstawową należy podczas konwersji przemnożyć przez wielokrotność liczby 10 w celu uzyskania czasu zapisanego w setkach nanosekund. Komentarze w pliku muszą rozpoczynać się znakiem #. Kropka oznacza koniec zapisu anotacji danego pliku.

2. Zebranie danych do analizy

Rozdział ten przedstawia szczegółowo proces przygotowania danych, z których będzie składał się korpus. Omawia on dokładnie, w jaki sposób zostały wygenerowane dane służące do przygotowania zestawów dla czytających oraz jakie są ich właściwości.

W rozdziale zostały też przedstawione dane, które były zbierane podczas procesu nagrywania oraz opisany system stworzony przez dyplomanta na potrzeby usprawnienia procesu zbierania danych dźwiękowych.

2.1. Typy danych

Na podstawie analizy materiałów dostępnych w sieci oraz własnych doświadczeń z telefoniczną weryfikacją rozmówcy zostały wybrane najczęściej używane typy danych, które następnie zostały wykorzystane do stworzenia korpusu. Te typy to: PESEL, data urodzenia, kod PIN, numer klienta, seria i numer dowodu osobistego.

2.2. Parametry typów danych

Aby jak najlepiej wytrenować program rozpoznający mowę (do czego może być też wykorzystany korpus), należało wygenerować dane treningowe jak najbardziej odpowiadające rzeczywistości. Dla numeru PESEL i daty urodzenia zostały określone z przedziału od 1 stycznia 1940 r. do 31 grudnia 2014 r. Dzięki temu można założyć, że wykorzystane dane mogą wytrenować system do prawidłowego działania przez najbliższe 18 lat. Ostatnie pięć cyfr numeru PESEL zostało wygenerowane zgodnie z obowiązującymi standardami, gdzie parzystość przedostatniej cyfry określa płeć, a ostatnia cyfra jest sumą kontrolną obliczoną z wzoru

$$a + 3b + 7c + 9d + e + 3f + 7g + 9h + i + 3j$$

gdzie:

litery od a do j oznaczają kolejne cyfry numeru PESEL.

Jako kod PIN został wygenerowany losowy ciąg czterech cyfr. Na 320 rekordów 40 powstało przez dwukrotne wklejenie listy 20 najczęściej występujących numerów PIN znalezionej na stronie internetowej datagenetics.com [23].

Numer klienta to ciąg 8 losowych cyfr. Numer dowodu osobistego zawiera tych cyfr o dwie mniej. Jako serię dowodu osobistego przyjąłem ciąg 3 liter, gdzie pierwszą zawsze jest A. Założenie to można przyjąć za prawdziwe, ponieważ maksymalna liczba dowodów osobistych wydanych w seriach rozpoczynających się na literę A wynosi 1 757

600 000 (zakładając 26 liter w alfabecie i 5 z 6 cyfr jako losowe, a szóstą, jako sumę kontrolną), co ponad czterdziestokrotnie przewyższa liczbę mieszkańców Polski [24].

Suma kontrolna dowodu osobistego jest obliczana w następujący sposób: serię dowodu zamienia się na wartości liczbowe wg wzoru [25]:

A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
10	11	12	13	14	15	16	17	18	19	20	21	22
N	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
23	24	25	26	27	28	29	30	31	32	33	34	35

Tabela 2.1 Wartości odpowiadające literom potrzebne do wyliczenia sumy kontrolnej

Aby sprawdzić poprawność numeru dowodu osobistego oblicza się sumę iloczynów cyfr/wartości odpowiadających literom oraz wag, które wynoszą kolejno 7 3 1 7 3 1 7 3. Suma tych iloczynów powinna po wykonaniu nań operacji modulo 10 dać wartość równą sumie kontrolnej.

2.3. Dopasowanie danych

Algorytm według którego zostały do siebie dopasowane zestawy danych musiał stworzyć zestawy jak najbardziej zróżnicowane między sobą, jeśli chodzi o częstotliwość występowania poszczególnych cyfr oraz liter. Algorytm, który został użyty, polegał na policzeniu częstości występowania każdej cyfry oraz litery i przedstawieniu ich w postaci liczb. Następnie dla każdej pary liczona była suma kwadratów liczb powstałych z sumowania współczynników przy tym samym znaku alfanumerycznym. Taka analiza danych zapewniała, że najmniejszy współczynnik (najlepsze dopasowanie) zostanie uzyskany dla danych, dla których rozkład występowania poszczególnych znaków będzie najbardziej zróżnicowany. Do wykonania obliczeń zostały użyte makra programu Microsoft Excel. Kod został przedstawiony w załączniku nr 1. W skrypcie został użyty algorytm dopasowania bąbelkowego. W rezultacie otrzymano dla każdej porcji danych pięć porcji, dla których współczynnik dopasowania był najmniejszy. Następne makro wybierało najlepiej dopasowane pary i odznaczało, które z porcji danych zostały już wykorzystane. Operacja ta zapobiegła duplikowaniu się danych w poszczególnych parach. Kod został przedstawiony w załączniku nr 2.

2.4. Metadane zbierane przy nagrywaniu

Jednym z wymagań postawionych korpusowi było zebranie odpowiednich metadanych do każdego nagrania, które mogą zostać później wykorzystane

do dokładniejszej analizy lub do lepszego rozpoznawania głosów na podstawie posiadanych informacji o nagraniach.

Metadane zbierane przy tworzeniu korpusu to:

- Płeć
- Przedział wiekowy
- Poziom szumu
- Rodzaj telefonu (stacjonarny/komórkowy)

Informacje o wieku osób nagrywanych, ze względu na ich prywatność, zostały ograniczone do informacji o przedziale wiekowym. Przyjęte zostały następujące przedziały wiekowe:

- < 20 lat
- 20-30 lat
- 31-40 lat
- 41-50 lat
- > 50 lat

Aby określić warunki, w jakich miało miejsce nagranie, przyjęto skalę pięciostopniową, określającą subiektywnie oceniony poziom szumu towarzyszącego nagrywanym słowom. Poziom „1" oznacza praktycznie brak szumów, poziom „2" szum ledwo odczuwalny, ale jednak dający się usłyszeć. Poziom „3" to szum wyraźnie słyszalny, ale w dość małym stopniu wpływający na pogorszenie rozpoznawalności poszczególnych wymawianych głosek. Przedostatnie dwa poziomy odpowiadają szumowi, który wpływa na odbiór nagrania oraz sprawia, że należy dokładniej przysłuchiwać się głoskom, by je rozpoznać. Przy poziomie szumu „5" osoba anotująca może mieć czasami problemy z rozpoznaniem głosek pomimo posiadania tekstu, który został przeczytany oraz wielokrotnego przesłuchania fragmentu nagrania. Ponadto, przy określeniu poziomu szumu wzięte zostały pod uwagę przypadki, gdy osoba nagrywająca mówiła cicho – wtedy też zostawała przyznana wyższa ocena wartości szumu.

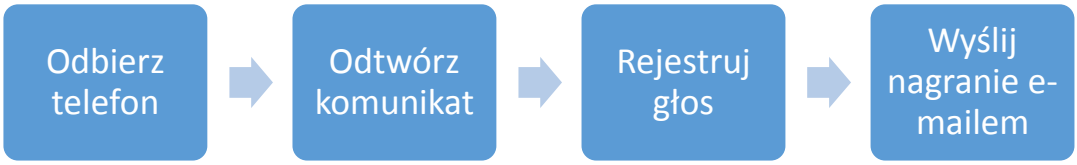
Przy określeniu płci nagrywającego użyto następującej zasady: kobieta została oznaczona literą „K" a mężczyzna literą „M". Rodzaj telefonu, z którego korzystała osoba nagrywająca został oznaczony jako „C" dla telefonu komórkowego i „S" dla telefonu stacjonarnego. Nie zostało wprowadzone bardziej szczegółowe rozróżnianie modelu telefonu.

2.5. System do zbierania nagrań

Zebranie materiału dźwiękowego zostało uproszczone przez dyplomanta dla zapewnienia komfortu nagrywającym, oraz aby zapobiec ewentualnym niejasnościom ze strony osób, których nagrane głosy miały znaleźć się w korpusie.

Do rejestrowania rozmów został wykorzystany system stworzony w oparciu o usługi jednego z operatorów telefonii VOIP (*ang. Voice over Internet Protocol*).

Algorytm działania systemu przedstawiał się następująco:



Rysunek 2.1 Algorytm działania systemu nagrywającego

Przed wykonaniem telefonu osoba nagrywająca się otrzymała zestaw danych, z którymi miała możliwość zapoznać się przed wykonaniem połączenia. Najczęściej w tym miejscu pojawiały się liczne wątpliwości dotyczące samego czytania. Dotyczyły one w największym stopniu sposobu odczytywania dat (czy należy je czytać cyfra po cyfrze, czy drugą z liczb należy czytać jako nazwę miesiąca czy też jako liczbę), jak również możliwości tworzenia złożeń cyfr (np. „Czy zamiast ‘jeden zero zero zero’ mogę przeczytać ‘tysiąc’?”). Aby nagrania były jak najbardziej zbliżone do rzeczywistości, dyplomant starał się nie narzucać nikomu sposobu odczytywania w/w danych. Spowodowało to wystąpienie w korpusie różnych wariantów ich odczytywania.

Po wyjaśnieniu wszystkich wątpliwości, osoba przystępowała do nagrywania. Sposób i miejsce nagrania nie było nigdy precyzyjnie określone, więc w większości przypadków nagrania były przeprowadzane w ustronnym miejscu, późnym wieczorem, żeby zapewnić jak najniższy poziom szumu oraz jak najlepszą jakość nagrania. Nie było to jednak regułą, ponieważ część sytuacji została umyślnie zaaranżowana tak, by nagranie zostało przeprowadzone w środowisku zaszumionym.

Po wybraniu podanego w instrukcji numeru telefonu, w słuchawce odtwarzany był nagrany wcześniej komunikat informujący, by rozpocząć czytanie po usłyszeniu sygnału dźwiękowego. Warto w tym miejscu również wspomnieć, że nagrania znacznie się między sobą różniły pod wieloma względami. Niektóre z nich były czytane bardzo

szybko i praktycznie z brakiem odstępów między frazami a niektóre powoli i bardzo dokładnie, z zachowaniem znacznych przerw. Z tego powodu najkrótsze nagrania trwały 40 sekund a najdłuższe 3 razy tyle. Dokładniejsza analiza statystyczna korpusu została przedstawiona w rozdziale 4.

Ważnym aspektem była też głośność wypowiadanych słów. W zależności od modelu telefonu oraz od samej osoby, która czytała dane, część nagrań była bardzo cicha, a część wręcz przesterowana. Mogło to też być spowodowane powszechnym przyzwyczajeniem do dość głośnego wypowiadania słów do telefonu.

Również poziom szumów w korpusie jest zróżnicowany. Jak zostało już wyżej wspomniane, część nagrań została przeprowadzona wieczorową porą w domach, a część w trakcie pracy czy podczas przebywania w otwartej przestrzeni miejskiej. Dzięki temu każde nagranie różni się od pozostałych. Zróżnicowanie pozwala również na stwierdzenie, że warunki, w których zostały poczynione nagrania, w dobry sposób odwzorowują realne warunki korzystania przez ludzi z telefonów.

2.6. Ciekawostki

Jedna osoba dostała do odczytania swoją datę urodzenia, a inna w numerze PESEL, który dostała do odczytania, odnalazła datę urodzenia własnego syna.

Skrótowiec „AGH” pojawił się w jednym z zestawów – nie został jednak przydzielony do studenta Akademii Górniczo Hutniczej.

Część tekstów została nagrana przez pracowników profesjonalnego Call Center (zestawy 110-117).

Najstarszy uczestnik nagrań miał w chwili nagrywania 74 lata.

3. Obróbka danych

3.1. Anotacja

Po zebraniu wymaganej liczby nagrań została przygotowana ich anotacja. Zgodnie z definicją tego pojęcia [26], zawiera ona również przygotowanie metadanych. Te zaś zostały omówione szczegółowo w rozdziale 2.4, więc w tym rozdziale zostanie dokładnie omówione jedynie przedstawienie zawartości nagrań za pomocą umownych oznaczeń tekstowych.

Jak zostało już wyżej wspomniane, aby zapewnić nagraniom naturalność oraz w realny sposób odwzorować rzeczywisty sposób wymowy, osobom nagrywającym się nie został narzucony sposób wymowy zarówno dat, jak i liczb. Pomimo oczywistych korzyści wynikających z tego zabiegu, jak np. duża różnorodność sposobów wymowy poszczególnych fraz, brak narzucenia reguł poskutkował również koniecznością zwiększenia nakładów pracy podczas znakowania tekstu.

Podczas procesu znakowania przyjęte zostały umowne zasady, dzięki którym dyplomant mógł z największą możliwą dokładnością opisać to, co zostało nagrane. Wielkie litery, które występowały w tekstach w polu „Dowód osobisty" zostały zapisane zgodnie ze sposobem ich wymawiania („a", „be", „ce", … „y" lub „igrek" itp.). Wszystkie pomyłki, zawahanie oraz niezidentyfikowane słowa powstałe przez problemy na łączu zostały oznaczone poprzez wzięcie ich w nawias oraz wpisanie w tym nawiasie dźwięku, który być może tam wystąpił. Taki zabieg jest przydatny przy detekcji tego rodzaju wypowiedzi.

Została również zwrócona uwaga na słowa, które w języku polskim mogą być wymawiane błędnie, takie jak, np. „jedynaście" zamiast „jedenaście" oraz „rozłanczam" zamiast „rozłączam". Znakowanie tego typu fraz zostało wykonane zgodnie z tekstem, który został wypowiedziany, niezależnie od jego poprawności językowej.

Nie wszystkie błędy w wypowiedziach zostały oznaczone. Przykładowo, liczbę 300 często można spotkać w nagraniach wymawianą jak „czysta" jednak obie wymowy są do siebie tak podobne, że nie można ze stuprocentową pewnością ocenić, jaki dźwięk wystąpił. Niektóre z dźwięków były też dźwiękami pośrednimi między wymową poprawną a błędną. W w/w przypadku taka wymowa wynikała z faktu, że wszystkie nagrania zostały wykonane przez osoby, które pochodziły lub aktualnie mieszkają

na terenach, na których używany jest dialekt małopolski, którego jedną z charakterystycznych cech jest używanie afrykat [27], czyli spółgłosek zwarto-szczelinowych [28].

3.2. Słowniki

Przygotowanie słowników było koniecznym elementem potrzebnym do analizy nagrań za pomocą systemu „AGH SARMATA". Przyjęto format słowników, w którym każde słowo znajduje się w osobnej linii, aby skorzystać z trybu seryjnego w systemie.

Do przygotowanie słowników został wykorzystany plik w formacie .csv zawierający znakowanie nagrań oraz skrypt Windows PowerShell, którego źródło zostało przedstawione w załączniku nr 4. Skrypt ten najpierw zapisuje wszystkie dane z pliku .csv do zmiennej (komenda Import-CSV). Następnie następuje obróbka tekstu, tj. usunięcie niepotrzebnych spacji (funkcja Trim) oraz zmiana wielkości liter na małe (funkcja ToLower) a następnie, za pomocą komendy Foreach-Object dopisuje każde niepuste (wyfiltrowane funkcją Where-Object) pole do końca pliku wynikowego. Plik ten jest osobny dla każdego nagrania.

Ostatnim etapem potrzebnym, by zapewnić zgodność z systemem SARMATA, była konwersja plików z formatu .mp3 na format .wav. Została ona wykonana za pomocą programu mpg123 działającego w systemie Linux Debian.

3.3. Analiza za pomocą programu SARMATA

W celu zoptymalizowania i automatyzacji części procesu przygotowywania korpusu dyplomant wykorzystał opisany w rozdziale 1.7 system rozpoznawania mowy SARMATA, który został do tego celu udostępniony dzięki uprzejmości Zespołu Przetwarzania Sygnałów.

Wykorzystanie SARMATY w zakresie niniejszej pracy inżynierskiej ograniczało się jedynie do automatycznej anotacji nagrań na podstawie plików w formacie .wav (16bit 16kHz PCM) oraz słowników (w plikach tekstowych w formacie UTF8 bez BOM).

Na wejście programu zostały podane również pliki zawierające plik służący do transkrypcji na zapis fonetyczny oraz baza wzorców. Z dostępnych dyplomantowi 2 baz została wybrana baza „complete_b00(…)", jako generująca lepsze wyniki (dokładniejsze wykrywanie ciszy). Została ona użyta pomimo dłuższego czasu przetwarzania nagrań.

Pozostałe parametry, z jakimi został uruchomiony program, to wyłączony pruning (algorytm drzew decyzyjnych służący do odrzucania najmniej wartościowych połączeń), brak Voice Activity Detection (system detekcji mowy używany m.in. w technologii VOIP) oraz włączona opcja śledzenia tylko jednej ścieżki (celem przyspieszenia obliczeń).

3.4. Konwersja do formatu .mlf

Pliki wyjściowe otrzymane z przetwarzania za pomocą SARMATY miały format tekstowy, jednak nie był to format zgodny z opisanym w rozdziale 1.8 formatem .mlf. Wycinek pliku tego formatu przedstawiał się następująco:

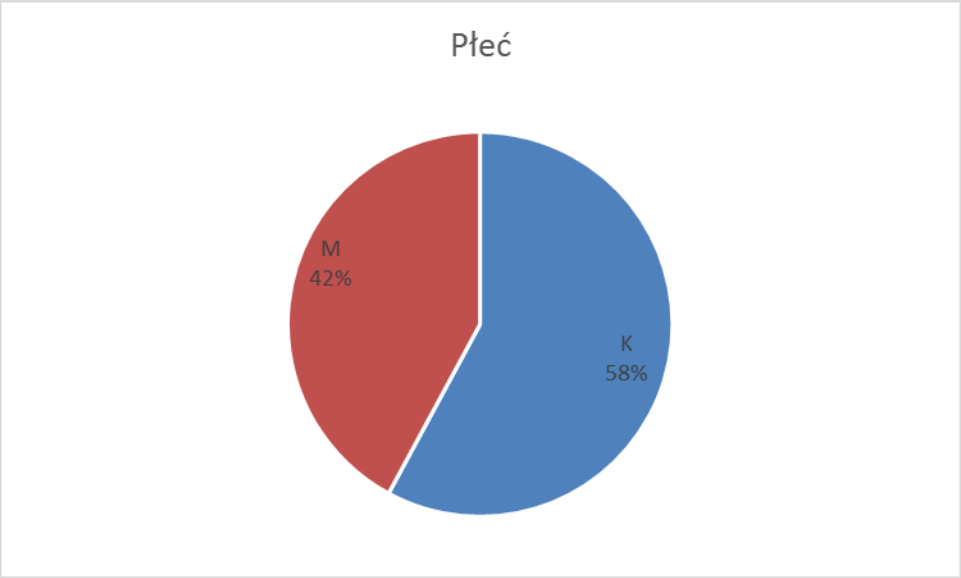
```
    1st Path:
1. sil [0 - 174] (0)
2. trzy [175 - 205] (0)
3. dwa [206 - 251] (0)
4. sil [252 - 293] (0)
Recognitions:
1.       sil trzy dwa sil (…)(-246445)
2.       sil trzy dwa sil (…) (-246445)
3.       sil trzy dwa sil (…)  (-246630)
4.       sil trzy dwa sil (…)(-246630)
```

Przetworzenie powyższego formatu na zgodny z formatem .mlf wymagało na początek zapisania do zmiennych niezbędnych danych (nazwy pliku, czasów rozpoczęcia i czasów zakończenia oraz słów). Następnie dane te należało zapisać do pliku .mlf pamiętając o odpowiedniej składni (m.in. nagłówek, nazwa pliku, znak kropki na końcu anotacji każdego pliku).

4. Analiza statystyczna korpusu

4.1. Osoby nagrywające się

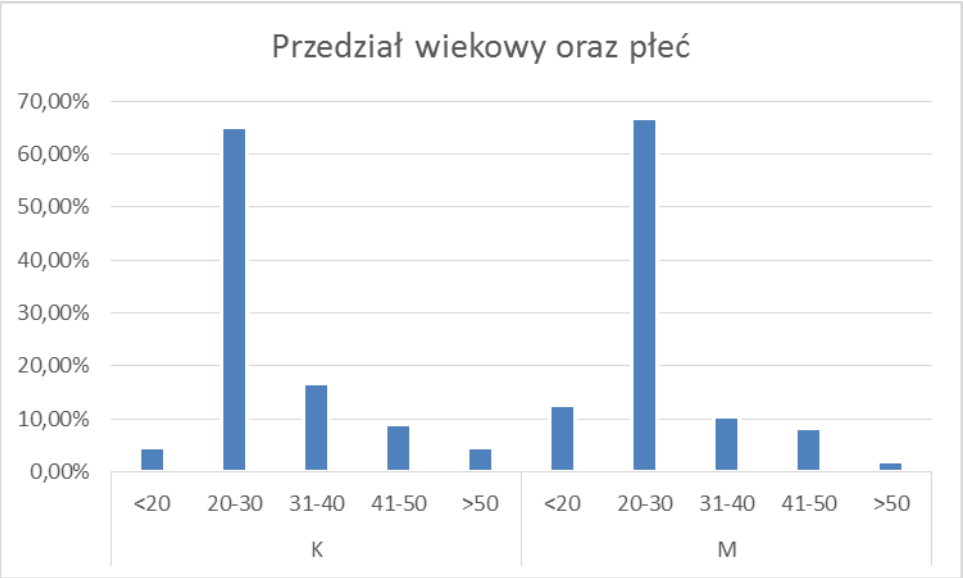
Do scharakteryzowania grupy, która uczestniczyła w nagraniach korpusu można użyć następujących wykresów:



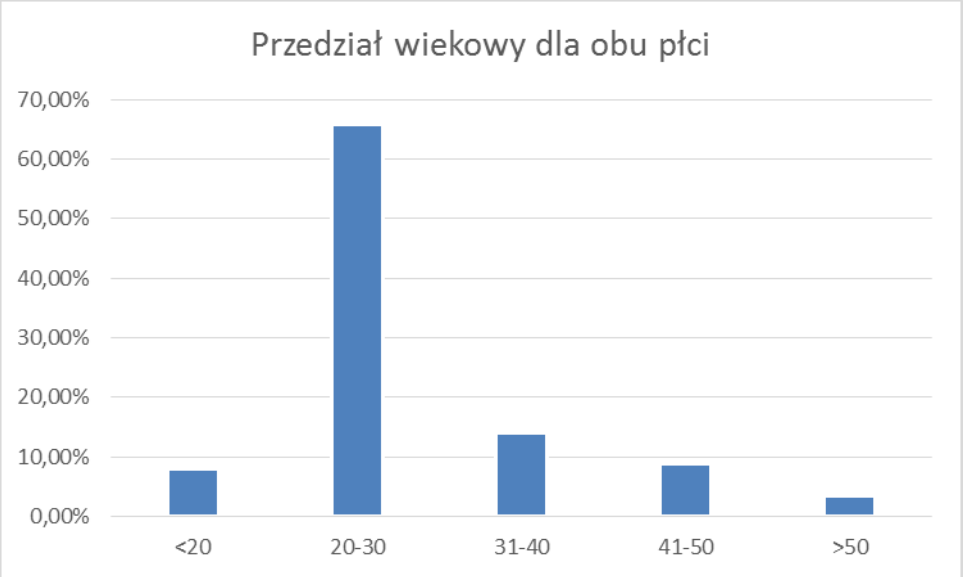
4.1 Wykres osób nagrywających się w zależności od płci

Powyższy wykres przedstawia jaki jest stosunek procentowy kobiet i mężczyzn w odniesieniu do całej grupy nagrywających się. W sumie grupę tę stanowi 66 kobiet oraz 48 mężczyzn.

W zależności od płci przedstawiono poniżej udział procentowy poszczególnych grup wiekowych w odniesieniu do wszystkich nagrywających się danej płci. Z dobrym przybliżeniem można przyjąć, że dla obojga płci udział grup wiekowych jest podobny. Największe różnice (ok. 10%) występują w grupach wiekowych poniżej 20 lat oraz w przedziale 31-40 lat. Ponadto, dla każdej płci wyraźnie przeważa grupa wiekowa 20-30 lat. Stanowi ona w obu przypadkach ponad 60% nagrywających się.



Wykres 4.2 Grupy wiekowe nagrywających z podziałem ze względu na płeć



Wykres 4.3 Wykres przedziałów wiekowych bez względu na płeć

Jak zostało już wcześniej wspomniane, z powodu niewielkich różnic w liczbie kobiet i mężczyzn, wykres nieuwzględniający płci wygląda bardzo podobnie do wykresu obrazującego przedziały wiekowe kobiet jak i do wykresu obrazującego przedziały wiekowe mężczyzn.

Jedynie niewielki odsetek połączeń został wykonany z telefonów stacjonarnych. Zapewne jest to spowodowane rosnącą w dalszym ciągu popularnością telefonów komórkowych oraz tym, że w dzisiejszych czasach już nawet najmłodsi posiadają takie urządzenia. Na wszystkie 114 nagrań jedynie 6 pochodziło z połączeń wykonanych telefonem stacjonarnym. Pozostałe 108 nagrań pochodzi z połączeń wykonanych telefonem komórkowym.

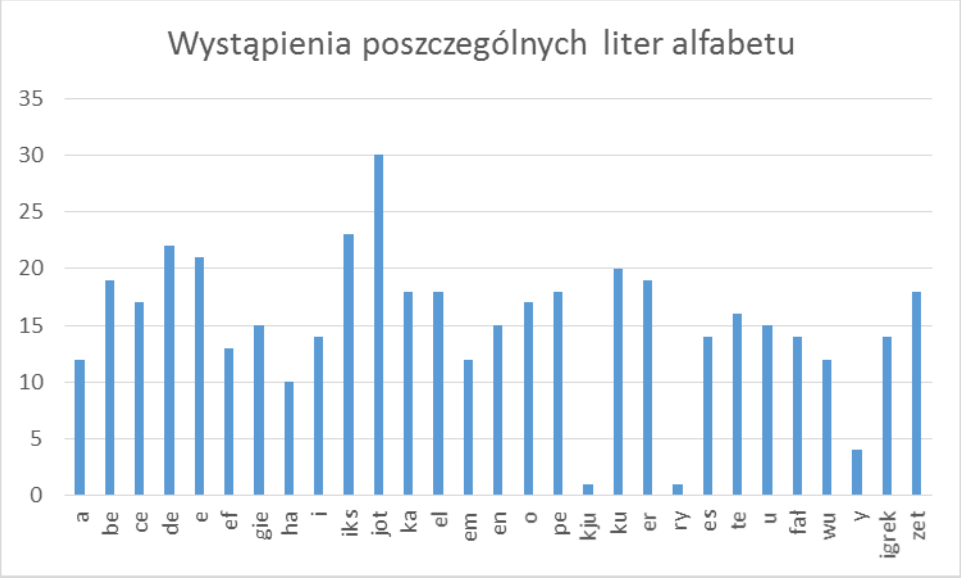
4.2. Słowa występujące w korpusie

Aby obliczyć częstotliwość występowania poszczególnych słów, zostało wykorzystane narzędzie dostępne na stronie internetowej [29].

W stworzonym korpusie występuje 198 unikalnych fraz, a ogółem zawiera on 8652 słowa. W obu tych liczbach zawarte są wszystkie słowa/frazy, które przy znakowaniu były traktowane jako oddzielny wyraz, czyli także litery, które były czytane podczas dyktowania pola „Dowód osobisty".

Pojedyncze litery występują w korpusie sumarycznie 670 razy, z czego sama litera „A" pojawia się 240 razy. Stanowi ona ponad jedną trzecią wystąpień wszystkich liter. Nie jest to zaskoczeniem, ponieważ jak zostało udowodnione w rozdziale 2.2, litera ta jest i jeszcze przez najbliższy czas będzie pierwszą literą serii każdego dowodu osobistego w Polsce.

Pomimo, że w korpusie nie wykorzystano polskich liter, występuje w nim 29 różnych fraz oznaczających litery. Alfabet łaciński zaś składa się jedynie z 26 znaków. 3 dodatkowe frazy spowodowane są różnym sposobem wymawiania „y" (jako „y" albo „igrek), „q" („ku" albo „kju") oraz „r" („er" albo „ry"). Są to jednakże jedynie pojedyncze przypadki.

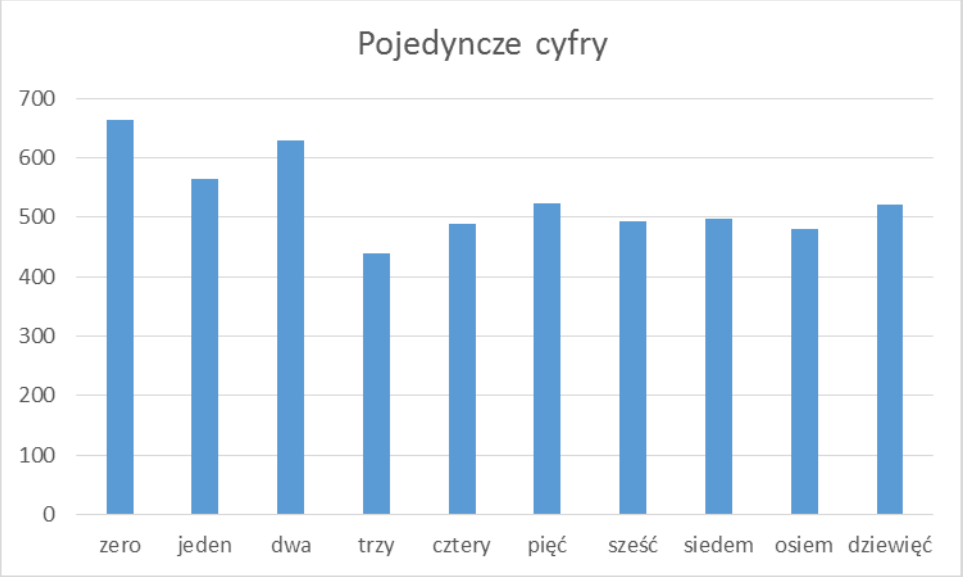


Wykres 4.4 Wystąpienia poszczególnych liter alfabetu

Powyższy wykres obrazuje niemal równą częstotliwość występowania poszczególnych liter alfabetu. Dla zachowania przejrzystości wartość wykresu odpowiadająca literze „a" została zmniejszona o 228 (tyle razy „a" występuje jako

pierwsza litera w serii „Dowodu osobistego”). 3 najniższe słupki obrazują alternatywne sposoby wymawiania „r”, „q” oraz „y”.

Poniższy wykres przedstawia zaś liczbę wystąpień poszczególnych cyfr (czytanych pojedynczo albo w złożeniach). Wyszczególniono na nim wystąpienia wyrazów w nieodmienionej formie. Sumarycznie poniższe słowa wystąpiły w korpusie 5303 razy.



Wykres 4.5 Wykres częstotliwości występowania poszczególnych cyfr

Słowem, które pojawia się najczęściej w całym korpusie jest słowo „zero”. Pojawia się ono 663 razy, co stanowi prawie 8% wszystkich wystąpień słów. Duża liczba wystąpień tego słowa jest spowodowana tym, że „zero” jest cyfrą, która, jeśli występuje z przodu, nie da się złożyć z inną cyfrą i przeczytać jako pary (wyjątkiem jest przypadek, kiedy „zero” jest po prostu opuszczane – zdarzyło się to w dwóch wypowiedziach).

Drugim, najczęściej pojawiającym się słowem, jest „dwa”. Warto zauważyć, że cyfra ta pojawia się jako pierwsza cyfra roku w dwudziestu procentach dat, które występują w polu „Data urodzenia” (zakres dat, jak zostało wspomniane w rozdziale 2.2 zawiera lata 1940-2014).

Wyrażenie „dwa tysiące” występuje więc 37 razy. 3 razy pojawia się słowo „dwutysięczny”, co sumarycznie daje 40 wystąpień. W pozostałych przypadkach liczba 2000 czytana była cyfra po cyfrze (ok. 10 wystąpień, co stanowi 20% wszystkich przypadków czytania liczby 2000).

Kolejną ważną cyfrą jest „jeden”. Jak można zobaczyć na stronie 28, liczba jej wystąpień w nagraniach tylko nieznacznie odbiega od wystąpień innych liczb (nie biorąc pod uwagę „dwa” oraz „zero”, które zostały omówione wcześniej).

Na pierwszy rzut oka może być to zastanawiające, ponieważ cyfra „1” jest nieodłącznym elementem niemal 80 procent dat występujących w korpusie. Biorąc jednak pod uwagę 185 wystąpień słowa „tysiąc”, można łatwo wytłumaczyć brak dużej przewagi wyrazu „jeden”. Jedynie w kilku przypadkach (na 228 dat występujących w korpusie), rok był czytany „cyfra po cyfrze”.

Zakończenie

Korpus mowy może zostać z powodzeniem wykorzystany do zwiększenia efektywności powtarzalnych zadań. Swoje biznesowe uzasadnienia zyskuje w momencie, gdy osoby, które np. muszą telefonicznie weryfikować dane rozmówcy, mogą zostać w tym etapie zastąpione przez odpowiedni automat. Może to pozwolić na szybszą i bardziej efektywną obsługę klienta, a przecież do takiej dąży każda firma posiadająca telefoniczną obsługę swoich kontrahentów.

Podsumowanie

Analizując przedstawione we wstępie wymagania dotyczące celu pracy oraz rezultat, którego omówieniem zajmuje się to wypracowanie, można stwierdzić, że zamierzony efekt został osiągnięty z nadwyżką ponad 20%.

Cały korpus został umieszczony na załączonej płycie CD. Zawiera ona zarówno pliki dźwiękowe, związane z nimi metadane, jak i otrzymane pliki .mlf. Korpus ten może zostać w przyszłości wykorzystany do automatyzacji wielu operacji wykonywanych za pomocą głosu oraz kombinacji cyfr. Przykładowe zastosowania mogą obejmować m.in. logowanie do różnych systemów, w których używa się kodu PIN. Warto również rozważyć połączenie powstałego korpusu z systemem weryfikacji biometrycznej opartym na analizie głosu, aby stworzyć metodę weryfikacji dwuetapowej.

Kolejne, przykładowe zastosowania, mogą się opierać na uproszczeniu procesu obsługi telefonicznej poprzez automatyczne pobranie od osoby obsługiwanej najczęściej potrzebnych danych. W przypadku obsługi technicznej, może to być np. wersja systemu operacyjnego, identyfikator oraz model sprzętu, którego dotyczy zgłoszenie. Autor pragnie również wyrazić nadzieję, że praca wykonana przez niego oraz inne osoby zaangażowane w przygotowanie korpusu, przyczyni się w pewnym stopniu do rozwoju prac związanych z przetwarzaniem języka polskiego oraz stanie się przydatnym narzędziem ułatwiającym codziennie wykonywane czynności.

Załączniki:

1. *Kod dopasowujący dane*

```
Sub dopasowanie2()
Dim i As Integer
Dim j As Integer
Dim k As Integer
Dim x As Integer
Dim y As Integer

Dim dopasowanie As Integer
Dim temp(1 To 321, 1 To 2) As Integer
Dim t As Integer
t = 1
For i = 2 To 321
For j = 2 To 321
If Worksheets("Arkusz1").Cells(i,6).Value =
Worksheets("Arkusz1").Cells(j, 6).Value Then
    dopasowanie = 0
    For k = 1 To 38
    x = Worksheets("Czest").Cells(i, k).Value
    y = Worksheets("Czest").Cells(j, k).Value
    dopasowanie = dopasowanie + ((x + y) * (x + y))

    Next k
    temp(j, 1) = dopasowanie
    temp(j, 2) = j
Else
    temp(j, 1) = 9999
    temp(j, 2) = j
End If

Next j
Call BubbleSort(temp)
Cells(i, 2).Value = temp(2, 1)
Cells(i, 3).Value = temp(3, 1)
Cells(i, 4).Value = temp(4, 1)
Cells(i, 5).Value = temp(5, 1)
Cells(i, 6).Value = temp(6, 1)
Cells(i, 7).Value = temp(2, 2)
```

```
    Cells(i, 8).Value = temp(3, 2)
    Cells(i, 9).Value = temp(4, 2)
    Cells(i, 10).Value = temp(5, 2)
    Cells(i, 11).Value = temp(6, 2)
    Cells(i, 12).Value = Worksheets("Arkusz1").Cells(i, 6)

    Next i
    End Sub
```

2. Kod eliminujący powtórzenia

```
    Sub wybierz()
    Worksheets("Arkusz1").Columns(7).ClearContents
    Dim i As Integer
    Dim j As Integer
    Dim x As Integer
    Dim y As Integer
    Dim k As Integer

    Dim wiersz As Integer
    k = 1
    x = 1
    y = 1

    Dim wartosc As Integer
    For j = 1 To 5
    For i = 2 To 321
    x = Worksheets("Arkusz2").Cells(i, 1).Value
    y = Worksheets("Arkusz2").Cells(i, j + 6).Value
    If Worksheets("Arkusz1").Cells(x, 7) <> "Wykorzystany" And
Worksheets("Arkusz1").Cells(y, 7) <> "Wykorzystany" Then
    Worksheets("Arkusz3").Cells(k, 1).Value =
Worksheets("Arkusz2").Cells(i, 1).Value
    Worksheets("Arkusz3").Cells(k, 2).Value =
Worksheets("Arkusz2").Cells(i, j + 6).Value
    Worksheets("Arkusz3").Cells(k, 3).Value =
Worksheets("Arkusz1").Cells(x, 6)
    wiersz = Worksheets("Arkusz2").Cells(i, 1).Value
    Worksheets("Arkusz1").Cells(wiersz, 7).Value = "Wykorzystany"
    wiersz = Worksheets("Arkusz2").Cells(i, j + 6).Value
    Worksheets("Arkusz1").Cells(wiersz, 7).Value = "Wykorzystany"
    k = k + 1
```

```
    End If
    Next i
    Next j
    End Sub
```

3. *Dane o rozmówcach – plik metadane.csv*

4. *Skrypt generujący słownik*

```
    $plik = "..\anotacja_spacje.csv"
    $dane = Import-Csv $plik -Delimiter ';'
    $dane| Foreach-Object {$_.PSObject.Properties | Foreach-
Object{$_.Value = $_.Value.Trim()}}
    $dane| Foreach-Object { $nazwa = $_.ID + ".txt";
$_.PSObject.Properties |Where-Object{($_.Name -notlike "ID") -and
($_.Value -ne "")}| Foreach-Object {$_.Value =
$_.Value.ToLower();$_.Value |Out-File $nazwa -Append}  }
```

5. *Plik opisujący występowanie słów – slowa.csv*

6. *Skrypt konwertujący plik wynikowy do formatu .mlf*

```
    $file=@()
    $wav=@()
    $j=0
    $mlf="zbiorczo.mlf"
    "#!MLF!#" |Out-File $mlf
    Get-ChildItem . -Filter *.txt |Foreach-Object{
    $file=$_.Name
    $wav=$file.Replace("txt","wav")

    "`"$wav`"" |Out-File $mlf -Append
    ""|Out-File $mlf -Append
    ""|Out-File $mlf -Append
    $wiersze=Get-Content .\$file |Where-Object {$_ -match "([0-9]*).
(\w*) \["}
    $kropka = @()
    $nawias1 = @()
    $kreska = @()
    $nawias2 = @()
    $czas1 = @()
    $czas2 = @()
    $tekst = @()
    $string = @()
```

```
    for ($i=0; $i -lt $wiersze.Length; $i++)
    {
    $kropka=$kropka+$wiersze[$i].indexof(".")
    $nawias1=$nawias1+$wiersze[$i].indexof("[")
    $kreska=$kreska+$wiersze[$i].indexof("-")
    $nawias2=$nawias2+$wiersze[$i].indexof("]")

    $czas1=$czas1+$wiersze[$i].Substring($nawias1[$i]+1,$kreska[$i]-
$nawias1[$i]-2)
    $czas2=$czas2+$wiersze[$i].Substring($kreska[$i]+2,$nawias2[$i]-
$kreska[$i]-2)
    $tekst=$tekst+$wiersze[$i].Substring($kropka[$i]+2,$nawias1[$i]-
$kropka[$i]-2)

    $string=$string + ($czas1[$i] + " " + $czas2[$i] + " " +
$tekst[$i])
    }
    $string |Out-File $mlf -Append
    "." |Out-File $mlf -Append
```

Bibliografia

[1] http://pl.wiktionary.org/wiki/korpus#korpus_.28j.C4.99zyk_polski.29. [Data uzyskania dostępu: 18 stycznia 2015].

[2] http://korpusy.net/component/option,com_glossary/id,8/. [Data uzyskania dostępu: 18 stycznia 2015].

[3] http://encyklopedia.pwn.pl/haslo/fonem;3901807.html. [Data uzyskania dostępu: 18 stycznia 2015].

[4] B. Ziółko i M. Ziółko, Przetwarzanie mowy, Kraków: Wydawnictwo AGH, 2011.

[5] L. R. Rabiner i B. H. Juang, Automatic Speech Recognition – A Brief History of the Technology Development, Atlanta: Georgia Institute of Technology, 2014.

[6] D. H. Klatt, „Review of the ARPA Speech Understanding Project," Massachusetts Institute of Technology, Cambridge, 1977.

[7] „The World's 10 most influential Languages," *Language Today*, pp. 12-18, 3 1997.

[8] Nationalencyklopedin "Världens 100 största språk 2007".

[9] http://www.speechlabs.pl/pl/project_ppbw. [Data uzyskania dostępu: 18 stycznia 2015].

[10] https://support.google.com/chrome/answer/1407892?hl=pl. [Data uzyskania dostępu: 18 stycznia 2015].

[11] http://youtu.be/4mtYb7a0loQ. [Data uzyskania dostępu: 18 stycznia 2015].

[12] http://www.dsp.agh.edu.pl/. [Data uzyskania dostępu: 18 stycznia 2015].

[13] S. Grocholewski, CORPORA - speech database for Polish diphones. Proceedings of Eurospeech., 1997.

[14] S. Grocholewski, „CORPORA - speech database for Polish diphones," w *Eurospeech*, Rhodos, Greece, 1997.

[15] G. Demenko, S. Grocholewski, K. Klessa, Lange M., M. Lange, D. Śledziński i N. Cylwik, „JURISDIC - Polish speech database for taking dictation of legal texts," w *Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[16] A. Przepiórkowski, M. Bańko, R. Górski i B. Lewandowska-Tomaszczyk, Narodowy Korpus Języka Polskiego, Warszawa: Wydawnictwo Naukowe PWN, 2012.

[17] B. Ziółko, T. Jadczyk, D. Skurzok i P. Żelasko, AGH Corpus of Polish Speech.

[18] http://www.dsp.agh.edu.pl/pl:resources:ortfon#.VLP3XFWG-I4. [Data uzyskania dostępu: 18 stycznia 2015].

[19] http://nkjp.pl/. [Data uzyskania dostępu: 15 stycznia 2015].

[20] http://www.dsp.agh.edu.pl/pl:resources:korpusav#.VLP_L1WG-QQ. [Data uzyskania dostępu: 18 stycznia 2015].

[21] http://www.dsp.agh.edu.pl/pl:resources:korpusemo#.VLP-6FWG-QQ. [Data uzyskania dostępu: 18 stycznia 2015].

[22] http://techmo.pl/index.php/voice-color/opis-produktu. [Data uzyskania dostępu: 18 stycznia 2015].

[23] „Data Genetics," wrzesień 2012. http://www.datagenetics.com/blog/september32012/. [Data uzyskania dostępu: 18 stycznia 2015].

[24] „Główny Urząd Statystyczny," 09 kwietnia 2013. http://stat.gov.pl/spisy-powszechne/nsp-2011/nsp-2011-wyniki/ludnosc-stan-i-struktura-demograficzno-spoleczna-nsp-2011,16,1.html. [Data uzyskania dostępu: 18 stycznia 2015].

[25] M. Kwiatek, „ Algorytmy i Struktury Danych," 03 marca 2008. http://www.algorytm.org/numery-identyfikacyjne/numer-dowodu-osobistego.html. [Data uzyskania dostępu: 18 stycznia 2015].

[26] http://korpusy.net/component/option,com_glossary/id,8/. [Data uzyskania dostępu: 18 stycznia 2015].

[27] http://sjp.pl/afrykata. [Data uzyskania dostępu: 18 stycznia 2015].

[28] H. Karaś, „Dialekty i gwary polskie,” http://www.dialektologia.uw.edu.pl/index.php?l1=leksykon&lid=589. [Data uzyskania dostępu: 18 stycznia 2015].

[29] http://rainbow.arch.scriptmania.com/tools/word_counter.html . [Data uzyskania dostępu: 11 stycznia 2015].