AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Inżynierii Mechanicznej i Robotyki

MAGISTERSKA PRACA DYPLOMOWA

ALEKSANDRA WYSZYŃSKA

ANALIZA KOMERCYJNYCH WDROŻEŃ SYSTEMU ROZPOZNAWANIA MOWY SARMATA

dr inż. Bartosz Ziółko
Promotor pracy

. . . . . .
Ocena, data,
podpis promotora

Kraków, rok 2012 / 2013

Kraków, dn. . . . . .

Imię i nazwisko: Aleksandra Wyszyńska
Nr albumu: 220067
Kierunek studiów: Inżynieria Akustyczna
Specjalność: Inżynieria dźwięku w mediach i kulturze

OŚWIADCZENIE AUTORA PRACY

Świadoma odpowiedzialności karnej za poświadczanie nieprawdy oświadczam, że niniejszą magisterską pracę dyplomową wykonałam osobiście i samodzielnie oraz nie korzystałam ze źródeł innych niż wymienione w pracy.
Jednocześnie oświadczam, że dokumentacja pracy nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz. U. z 2006 r. Nr 90 poz. 631 z późniejszymi zmianami) oraz dóbr osobistych chronionych prawem cywilnym. Nie zawiera ona również danych i informacji, które uzyskałam w sposób niedozwolony. Wersja dokumentacji dołączona przeze mnie na nośniku elektronicznym jest w pełni zgodna z wydrukiem przedstawionym do recenzji.
Zaświadczam także, że niniejsza magisterska praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów wyższej uczelni lub tytułów zawodowych.

. . . . . .

podpis

Kraków, dn. . . . . .

Imię i nazwisko: Aleksandra Wyszyńska

Adres korespondencyjny: ul. Mioceńska 1/43, 97-400 Bełchatów

Temat magisterskiej pracy dyplomowej: Analiza komercyjnych wdrożeń systemu rozpoznawania mowy SARMATA

Nr albumu: 220067

Kierunek studiów: Inżynieria Akustyczna

Specjalność: Inżynieria dźwięku w mediach i kulturze

OŚWIADCZENIE

Niniejszym oświadczam, że zachowując moje prawa autorskie, udzielam Akademii Górniczo-Hutniczej im. S. Staszica w Krakowie nieograniczonej w czasie nieodpłatnej licencji niewyłącznej do korzystania z przedstawionej dokumentacji magisterskiej pracy dyplomowej, w zakresie publicznego udostępniania i rozpowszechniania w wersji drukowanej i elektronicznej. [1]

Kraków, dn. . . . . .

data, podpis

[1] Na podstawie Ustawy z dnia 27 lipca 2005 r. Prawo o szkolnictwie wyższym (Dz.U. 2005 nr 164 poz. 1365) Art. 239. oraz Ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (Dz.U. z 2000 r. Nr 80, poz. 904, z późn. zm.) Art. 15a. „Uczelni w rozumieniu przepisów o szkolnictwie wyższym przysługuje pierwszeństwo w opublikowaniu pracy dyplomowej studenta. Jeżeli uczelnia nie opublikowała pracy dyplomowej w ciągu 6 miesięcy od jej obrony, student, który ją przygotował, może ją opublikować, chyba że praca dyplomowa jest częścią utworu zbiorowego."

Kraków, ...

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Wydział Inżynierii Mechanicznej i Robotyki

TEMATYKA MAGISTERSKIEJ PRACY DYPLOMOWEJ

dla studenta II roku studiów stacjonarnych

Aleksandra Wyszyńska

TEMAT MAGISTERSKIE PRACY DYPLOMOWEJ:

Analiza komercyjnych wdrożeń systemu rozpoznawania mowy SARMATA

Miejsce praktyki dyplomowej: nie dotyczy

Promotor pracy: dr inż. Bartosz Ziółko

Recenzent pracy: . . . . . .

. . . . . .

podpis dziekana

PROGRAM PRACY I PRAKTYKI DYPLOMOWEJ:

1. Omówienie tematu pracy i sposobu realizacji z promotorem.
2. Zebranie i opracowanie literatury dotyczącej tematu pracy.
3. Zebranie i opracowanie wyników badań.
4. Analiza wyników badań, ich omówienie i zatwierdzenie przez promotora.
5. Opracowania redakcyjne.

Kraków, ...

data i podpis dyplomanta

Termin złożenia do dziekanatu: . . . . . .

. . . . . .

podpis promotora

Kraków, ...

Akademia Górniczo-Hutnicza im. Stanisława Staszica

Wydział Inżynierii Mechanicznej i Robotyki

Kierunek studiów: Inżynieria Akustyczna

Specjalność: Inżynieria dźwięku w mediach i kulturze

Aleksandra Wyszyńska

Magisterska praca dyplomowa

Analiza komercyjnych wdrożeń systemu rozpoznawania mowy SARMATA

Promotor: dr inż. Bartosz Ziółko

STRESZCZENIE

Niniejsza praca dyplomowa odnosi się do zagadnień związanych z rozpoznawaniem mowy. Opracowano i przetestowano bazy GMM (ang. Gaussian Mixture Models) o różnych parametrach służące do stworzenia statystycznych modeli języka. W tym celu stworzono zestawy testowe oraz słowniki zawierające zapis fonetyczny zestawów. Przetestowano także program Anotator2.0 służący do segmentacji mowy. Opisano stanowisko VoIP, które w przyszłości posłuży do stworzenia korpusu rozmów telefonicznych.

Krakow, the .....

AGH University of Science and Technology

Faculty of Mechanical Engineering and Robotics

Field of Study: Acoustic Engineering

Specialisations: Sound engineering in media and culture

Aleksandra Wyszyńska

Master Diploma Thesis

Analysis of commercial applications of SARMATA speech recognition system

Supervisor: Bartosz Ziółko Ph.D

ABSTRACT

This thesis describes issues related to automatic speech recognition. Gaussian Mixture Models with different parameters were prepared and tested. GMM models were used for making statistical language models. For this purpose testing data and dictionaries containing phonetic notation of the data were created. Program Anontator 2.0, used in segmentation of speech records, was also tested. Finally, the VoIP terminal configuration is described, which is thought to be a part of system creating new telephone conversation corpus.

Składam serdecznie podziękowania dr. Bartoszowi Ziółce

Kochanym Rodzicom

Spis treści

1. Wprowadzenie

Podstawowym i najbardziej naturalnym sposobem komunikacji międzyludzkiej jest mowa. Oczywistym więc jest, że człowiek chciałby się w ten sposób komunikować także z maszynami. Właśnie dlatego powstają systemy rozpoznawania mowy. W niniejszej pracy opisano testy jednego z systemów rozpoznawania mowy, SARMATA, który jest rozwijany przez Zespół Przetwarzania Sygnałów DSP AGH.

W rozdziale pierwszym opisano zagadnienia związane z rozpoznawaniem mowy, których dotyczyć będzie ta praca. Omówiono historię systemów rozpoznawania mowy, czym są korpusy mowy, czego dotyczy pojęcie segmentacji oraz klasyfikacji.

W rozdziale drugim opisano prace związane z tworzeniem baz GMM, czyli statystycznego modelu języka. Przedstawiono jak stworzono zestawy testowe, słowniki do tych zestawów omówiono zasady fonetyki, które potrzebne były do stworzenia słowników oraz zaprezentowano wyniki testu przeprowadzonego na różnych bazach GMM.

W rozdziale trzecim opisano test programu Anotator2.0. Program ma być pomocny w segmentacji nagrań oraz przyspieszać pracę osoby tworzącej pliki mlf. Przetestowano jego działanie i opisano wyniki.

W rozdziale piątym opisano prace związane ze stworzeniem stanowiska VoIP, które posłuży do stworzenia korpusu rozmów telefonicznych. Opisano czym jest VoIP (ang. *Voice over Internet Protocol*), jego wady i zalety, oraz jak wyglądało samo przygotowanie stanowiska.

2. Rozpoznawanie mowy

W tym rozdziale opisano elementy systemów rozpoznawania mowy, którymi zajmowano się podczas tworzenia tej pracy dyplomowej.

2.1. Rys historyczny

Jedną z pierwszych osób, których prace bardzo przyczyniły się do powstania systemów rozpoznawania mowy był Aleksander Graham Bell. Próbował on stworzyć urządzenie, któremu można dyktować tekst. Zadanie to nie powiodło się, jednak wynik jego prac, czyli telefon (a dokładniej wchodzący w jego skład mikrofon) był elementem umożliwiającym dalsze badania[2] [7].

Pierwszym systemem rozpoznawania mowy możemy nazwać psa zabawkę Radio Rex powstałą w 1920 roku. Figurka psa, która znajdowała się w budzie, za każdym razem gdy wymówiliśmy imię psa ”Rex” wyskakiwała z budy. Działo się to dzięki zjawisku rezonansu akustycznego- płytka do której przymocowana była figurka reagowała obrotem na drgania o częstotliwości 500Hz, czyli tym odpowiadającym głosce ”e” [2] [7].

W latach 30-tych XX wieku Stevens i Newman zdefiniowali melową skale częstotliwości, która jest wykorzystywana do dziś w rozpoznawaniu mowy [7].

Dużym postępem było opracowanie przez naukowców z Bell Labs w 1952 systemu rozpoznawania cyfr izolowanych. Angielskie cyfry były rozpoznawane z błędem mniejszym niż 2%, ale jedynie gdy układ ust mówiącego względem mikrofonu był taki sam w czasie ustalania parametrów głosu i w czasie testów.

W latach 60-tych XX wieku opracowano algorytm szybkiej transformacji Fouriera (ang. Fast Fourier Transform), która skróciła znacznie obliczenia pozwalające na analizę widma, oraz niejawne modele Markowa (ang. Hidden Markov Model- HMM) stosowane do modelowania mowy. HMM wykorzystuje prawdopodobieństwo wystąpienia głosek przy zaobser-

wowanych parametrach mowy. Zarówno FFT jak i HMM do dzisiaj są podstawami systemów rozpoznawania mowy [2][7].

W latach 90-tych wprowadzono pierwsze dostępne dla przeciętnego użytkownika systemy ASR (ang. Automatic Speech Recognition- systemy rozpoznawania mowy) takie jak Dragon, czy IBM ViaVoice [7].

Obecnie systemy ASR możemy podzielić na dwa podstawowe typy: systemy rozpoznawania słów izolowanych z ograniczonym słownikiem(IWRS, ang Isolated Word Recognition Systems) oraz system rozpoznawania mowy ciągłej i swobodnej z bardzo dużym słownikiem(LVCSR, ang. Large Vocabulary Continuous Speech Recognition). Systemy IWRS osiągają wysoką skuteczność, jednak nie są to systemy pozwalające swobodnie komunikować się z maszyną. Systemy LVCSR są systemami dużo bardziej skomplikowanymi i rozbudowanymi. Pomiędzy tymi dwoma rozwiązaniami jest wiele rozwiązań pośrednich (np. systemy z ograniczonym słownikiem) [7].

2.2. Korpusy mowy

Systemy rozpoznawania mowy do stworzenia statystycznych modeli języka (procesu nazywanego szkoleniem lub treningiem) potrzebują dużej ilości danych, na których będą się opierały. Zbiorem takich danych językowych mogą być nagrania, teksty, strony internetowe etc. Im więcej danych dostarczymy systemowi, tym większą skutecznością będzie się wykazywał [2]. W tym celu tworzone są zasoby zawierające niejednokrotnie więcej danych niż jedynie nagrania i ich transkrypcje (np. podział na wypowiedzi, analizę morfologiczną słów, wyodrębnione proste frazy itp.). Duży zbiór takich danych nazywamy korpusem mowy [2]. Są one trudne do stworzenia, ze względu właśnie na dodatkowe dane, które zawierają, tworzone najczęściej ręcznie, w związku z czym wymagają czasu i cierpliwości, a co za tym idzie dużych nakładów pieniężnych.

Ze względu na bogactwo językowe bardzo trudno stworzyć korpus zawierający wszystkie pojęcia i słowa. Najczęściej tworzy się je więc w oparciu o temat, do jakiego będzie wykorzystywany ASR, co znacznie pozwala ograniczyć słownik. Mogą one także opierać się na mowie spontanicznej (nagrania rozmów) lub na tekście pisanym (osoby czytające tekst). Pierwszy rodzaj korpusów jest bardziej pożądany, ze względu na to, że język w mowie swobodnej różni się znacząco od tekstu pisanego. Jednakże dużo łatwiej uzyskać nagrania dobrej

jakości prosząc osobę o przeczytanie tekstu w warunkach studyjnych i jest to o wiele tańsze i szybsze.

W rozdziale 5 opisano przygotowanie stanowiska VoIP, które będzie służyło do nagrywania rozmów telefonicznych i pozyskania nagrań do korpusu mowy spontanicznej.

W tej pracy wykorzystano dwa korpusy mowy polskiej: korpus polskich rozmów telefonicznych LUNA oraz Korpus GlobalPhone.

2.2.1. Korpus polskich rozmów telefonicznych LUNA

Korpus polskich rozmów telefonicznych LUNA powstał w ramach projektu LUNA (ang. *spoken Language UNderstanding In MultilinguAl Communications Sysytem*). Celem tego projektu było opracowanie narzędzi do dostosowywania oprogramowania rozpoznawania mowy do nowego języka lub nowej tematyki dialogu. Projekt skupiał się na zagadnieniach związanych z rozumieniem mowy w językach francuskim, włoskim oraz polskim [10].

W skład polskiego korpusu rozmów telefonicznych LUNA wchodzą dwa korpusy mowy: korpus rozmów człowieka z człowiekiem oraz rozmów człowieka z komputerem. Każdy z tych korpusów zawiera 500 dialogów, które zostały zanotowane na różnych poziomach. Parametry nagrań to: częstotliwość próbkowania- 16 kHz, rozdzielczość- 16 bitów, system mono. W ramach projektu inżynierskiego *Rozwinięcie korpusu polskich rozmów telefonicznych LUNA* został on poszerzony o pliki mlf (2.3.1).

Korpus mowy Luna jest wyjątkowym korpusem, ze względu na to, iż zawiera zapis spontanicznych rozmów, a nie sztywno zarysowanych wypowiedzi.

2.2.2. Korpus mowy GlobalPhone

GlobalPhone to korpus mowy, w skład którego wchodzą korpusy 20 języków takich jak: arabski, bułgarski, chiński (mandaryński i szanghajski), chorwacki, czeski, angielski, francuski, niemiecki, japoński, koreański, polski, portugalskim rosyjski, hiszpański, szwedzki, tamilski, tajski, turecki i wietnamski. Korpus zawiera ponad 400 godzin nagrań mowy ponad 1900 mówców. W każdym z języków około 100 mówców czyta około 100 zdań. Zdania pochodzą z gazet oraz internetu. Parametry nagrań to: 16bit i 16kHz w systemie mono [12].

2.3. Segmentacja

System rozpoznawania mowy musi z sygnału uzyskać i przetworzyć wiele informacji. Pojedyncza próbka niesie niewystarczającą ilość takich informacji, w związku z czym projektant systemu musi uporać się z segmentacją sygnału, czyli podzieleniem sygnału na ramki odpowiedniej długości. Podczas tego procesu należy zwrócić uwagę na efekty brzegowe występujące podczas ramkowania sygnału. Aby zniwelować zniekształcenia stosuje się ramki o większej długości (im krótsza ramka tym wpływ zniekształceń jest większy) oraz do stosuje się okno o wąskim widmie (najczęściej jest to okno Hamminga)[7].

Wyróżniamy dwa podstawowe typy segmentacji:

– segmentację równomierną;

– segmentację nierównomierną.

"Segmentacja równomierna jest najprostszym i najczęściej stosowanym typem segmentacji" [7]. Gdy ten rodzaj segmentacji stosujemy do sygnału mowy najczęściej używamy ramek długości 20 ms, ponieważ jest to średni czas trwania najkrótszych fonemów. W praktyce stosuje się ramki o długości $N = 2^k$, np. N=256, co przy częstotliwości próbkowania $f_s = 16$ kHz daje długość równą 16 ms [7][9]. W segmentacji równomiernej stosuje się ramkowanie z zakładką, aby zwiększyć rozdzielczość analizy.

Segmentacja nierównomierna ma na celu podzielić sygnał ze względu na jego treść (najczęściej wyodrębnione zostaną fonemy, difony lub trifony). Gdy sygnał ma zostać podzielony na większe partie (np. słowa) najczęściej dokonuje tego już człowiek.

2.3.1. Pliki mlf

Plik mlf jest to plik zawierający czas początku i końca każdego słowa w nagraniu. Dzięki tym plikom można wykonać trening systemu rozpoznawania mowy i wyliczyć parametry dla modeli HMM, GMM czy algorytmu kNN [2][16].

W rozdziale 4 omówiono program Anotator, który służy do rozwijania korpusów mowy- pozwala na tworzenie plików mlf. Program został wzbogacony w ASR, aby ułatwić to żmudne zadanie. Następnie dzięki tym plikom program sam będzie wstanie z mniejszych partii sygnału(ze słów) dokonać segmentacji sygnału na fonemy. Rozdział 4 omawia jak sprawdzono sprawność ASR załączonego do programu, na podstawie nagrań z dwóch różnych korpusów mowy: LUNA oraz GlobalPhone.

2.4. Metody klasyfikacja sygnału

"Klasyfikacja to proces przyporządkowania obiektów (np. fragmentów sygnału mowy) do pewnych klas (np. konkretnych fonemów)"[7]. W tym podrozdziale zostanie opisany klasyfikator k-NN oraz GMM (ang. Gaussian Mixture Models), które zostały wykorzystane w dalszej części pracy 3.

2.4.1. Klasyfikator k-NN

Klasyfikator k-NN (k- Najbliższych Sąsiadów, ang. k-Nearest Neighbors) to jedna z metod minimalno odległościowych, które polegają na wybraniu klasy, do której należy obiekt leżący najbliżej (według przyjętej metryki- euklidesowej, Czebyszewa, Minkowskiego etc) rozpoznawanemu obiektowi z ciągu uczącego. Klasyfikator k-NN sprawdza, "w jakich klasach średnia odległość klasyfikowanego wektora, od k elementów klas, jest najmniejsza"[7].

Numer klasy dla klasyfikatora k-NN:

$$i^* = \arg\min_i(\overline{\delta_i}).$$

Średnia odległość dla tej klasy:

$$\overline{\delta_i} = \min_{X_i^* \subset X_i} (\sum_{x_n \in X_i^*} \delta(x_n, x), |X_i^*| = k,$$

k najbliższych wzorców od wektora x jest najmniejsza. X jest to ciąg wektorów [7].

Klasyfikator k-NN jest jednym z najczęściej stosowanych dzięki łatwej implementacji i dużej skuteczności.

2.4.2. Złożone modele Gaussa

Złożone modele Gaussa (Gaussian Mixture Models - gmm) to funkcja gęstości prawdopodobieństwa reprezentowanego jako suma gaussowskich gęstości składowych [13]. Modele GMM są używane jako parametryczne modele rozkładu prawdopodobieństwa cech ciągłych w systemach biometrycznych.

Złożone modele Gaussa to suma M składowych gaussowskich gęstości jak w równaniu:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \sum_i)$$

gdzie x to D-wymiarowy wektor ciągłych wartości (np. pomiar cech), w_i, $i = 1, ..., M$, są wagą modelu, zaś $g(x|\mu_i, \sum_i)$, $i = 1, ..., M$ to gaussowskie gęstości składowe. Każda gęstość składowa to D-wymiarowy rozkład Gaussa:

$$g(x|\mu_i, \sum_i) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)'\sum_i^{-1}(x-\mu_i)\right\},$$

z głównym wektorem μ_i i macierzą kowariancji $\sum_i$. Wagi modelu muszą spełniać warunek $\sum_{i=1}^{M} w_i = 1$ [13].

Wektor główny, macierz kowariancji oraz waga modeli gęstości wszystkich składowych są głównymi parametrami złożonych modeli Gaussa:

$$\lambda = \{w_i, \mu_i, \sum_i\} \quad i = 1, ..., M.$$

Istnieje kilka wariantów dla parametrów z powyższego równania. Macierze kowariancji $\sum_i$ mogą być diagonalne lub nie. Parametry mogą być ze sobą powiązane, udostępniane pomiędzy składowymi np. o wspólnej kowariancji dla wszystkich składników. Wybór konfiguracji modelu (liczba elementów, czy macierz kowariancji jest diagonalna) są często określane przez ilość dostępnych danych do oszacowania parametrów modelu GMM oraz jak model jest stosowany w danym systemie [13].

3. Generowanie i test baz GMM dla systemu rozpoznawania mowy SARMATA

Jednym z zadań przeprowadzonych w ramach tej pracy magisterskiej była pomoc w obliczeniach złożonych modeli Gaussa dla systemu SARMATA. Trening mlf prowadzący do wygenerowania baz GMM wykonano na plikach mlf powstałych w projekcie inżynierskim *Rozwinięcie korpusu polskich rozmów telefonicznych LUNA*.

3.1. Trening mlf systemu rozpoznawania mowy SARMATA

3.1.1. Trening mlf oraz wyliczenie modeli GMM

Wykonano trening mlf systemu SARMATA i w ten sposób wygenerowano pliki * _patterns.dat oraz *_patterns_CUMULATED.dat zawierające parametry KNN dla fonemów występujących w nagraniach. Trening mlf wykonywano na 404 nagraniach pochodzących z korpusu polskich rozmów telefonicznych LUNA (więcej w rozdziale 2.2.1). Następnie dzięki powstałym plikom wyliczono parametry modelu GMM dla każdego fonemu dla różnej ilości komponentów (2, 3, 4, 5, 6, 8, 10, 12, 15, 18, 20). Powstałe w ten sposób pliki zawierały parametry μ_i, $\sum_i$ oraz wagę modeli i posłużyły do stworzenia baz wzorców GMM dla odpowiedniej ilości komponentów. Każda z baz składała się z modeli GMM dla każdego z 37 fonemów występującego w języku polskim 3.1.

3.1.2. Przygotowanie danych testowych

Następnie aby wykonać testy wygenerowanych modeli należało stworzyć foldery zawierające nagrania testowe. W tym celu stworzono 20 zestawów, po 10 dla nagrań pochodzących z plików:

1. Na których wykonano trening;

2. Na których nie wykonywano treningu.

W każdym z 20 zestawów znalazło się 30 plików *.wav zawierających spójne fragmenty (bez ciszy w środku) co najmniej jednego słowa (częściej sekwencje dwóch lub trzech słów). Pliki pochodzące z nagrań typu (1) oraz (2) nie były w folderach wymieszane. Pliki *.wav musiały być plikami mono o częstotliwości próbkowania 16 kHz oraz rozdzielczości 16 bit.

Do każdego z plików należało stworzyć plik *.txt zawierający transkrypcję zawartości oraz plik *.mlf, których zawartość wyglądała następująco:

#!MLF!#

“nazwa_pliku_01.wav“

0 123456789000000 cała_zawartość_wypowiedzi_tego_pliku_zgodna_ze_słownikiem

.

“nazwa_pliku_02.wav“

0 123456789000000 cała_zawartość_wypowiedzi_tego_pliku_zgodna_ze_słownikiem

.

.....

“nazwa_pliku_30.wav“

0 123456789000000 cała_zawartość_wypowiedzi_tego_pliku_zgodna_ze_słownikiem

.

Przykład początku pliku z folderu “1“:

#!MLF!#

“1_2007-03-13_12_58_32_01.wav“

0 5050000 dzień_dobry

.

“1_2007-03-13_12_58_32_02.wav“

10000 9090000 ja_chcę_zapytać

.

“1_2007-03-13_12_58_32_03.wav“

0 19910000 żeby_być_na_Dworcu_Zachodnim

.

“1_2007-03-13_12_58_32_04.wav“

20000 7180000 ale_z_pętli

.

"1_2007-03-13_12_58_32_05.wav"

30000 8520000 tej_na_Sadybie

.

"1_2007-03-13_12_58_32_06.wav"

10000 9690000 jak_mam_jechać

.

"1_2007-03-13_12_58_32_07.wav"

30000 12890000 jak_mam_jechać

.

...

3.1.3. Przygotowanie słowników dla plików testowych

Dla każdego z zestawów testowych stworzono plik dict_test.txt zawierający transkrypcje wszystkich nagrań znajdujących się w danym zestawie. Następnie przy pomocy programu SARMATA, wygenerowano pliki dictionary.txt- słowniki zawierające transkrypcje fonetyczne tekstu przedstawione alfabetem fonetycznym AGH pokazanym w tabeli 3.1 wraz z częstością występowania fonemów według różnych badań. Słowniki wymagały ręcznej korekty ze względu na zawiłości polskiej fonetyki.

Pracując nad transkrypcją fonetyczną trzeba zwrócić uwagę na dwa ważne zjawiska jakimi są koartykulacja oraz wynikające z niej upodobnienia.

Koartykulacja

Koartykulacja to ruchy narządów mowy przygotowujących się do wyartykułowania następnej głoski. Efekt akustyczny koartykulacji nazywamy przejściem tranzjentowym [15]. Przez ten proces następują upodobnienia, czyli pod wieloma względami jedna głoska staje się podobna do głoski z nią sąsiadującej [5].

Upodobnienia

Upodobnienia dzielimy na wewnątrzwyrazowe i międzywyrazowe [15][5], oraz ze względu na miejsce artykulacji, pod względem artykulacji oraz pod względem dźwięczności.

Upodobnienia pod względem miejsca artykulacji mają różny stopień nasilenia. Czasem poprzez te upodobnienia dochodzi do uproszczeń i powstają formy błędne, np zdanie [0terXie58i 0y 0terna58ie] można przeczytać jako *"43,14"* lub *"40 czy 14"* [5].

Tablica 3.1: Wybrane alfabety fonetyczne dla jęz. polskiego [2]

SAMPA	Grocholewski	AGH	ortogr.	fonetycz.	% [1]	% [4]
#				#	17.10	4.7
e	e	e	test	test	8.11	10.6
a	a	a	pat	pat	7.91	9.7
o	o	o	pot	pot	7.52	8.0
j	j	j	jak	jak	3.46	4.4
n	n	n	nasz	naS	3.39	4.0
t	t	t	test	test	3.39	4.8
i	i	i	PIT	pit	3.39	3.4
I	y	y	typ	tIp	3.37	3.8
r	r	r	ryk	rIk	2.98	3.2
v	w	v	wilk	vilk	2.89	2.9
m	m	m	mysz	mIS	2.76	3.2
p	p	p	pik	pik	2.65	3.0
u	u	u	puk	puk	2.62	2.8
s	s	s	syk	sIk	2.54	2.8
d	d	d	dym	dIm	2.18	2.1
k	k	k	kit	kit	2.09	2.5
w	l_	w	łyk	wIk	2.05	1.8
n’	ni	3	koń	kon’	1.97	2.4
l	l	l	luk	luk	1.92	1.9
z	z	z	zbir	zbir	1.67	1.5
g	g	g	gen	gen	1.38	1.3
b	b	b	bit	bit	1.34	1.5
S	sz	S	szyk	SIk	1.32	1.9
f	f	f	fan	fan	1.19	1.3
s’	si	5	świt	s’vit	1.16	1.6
Z	rz	Z	żyto	ZIto	1.06	1.3
tˆs	c	7	cyk	tˆsIk	1.06	1.2
x	h	x	hymn	xImn	1.01	1.0
tˆS	cz	0	czyn	tˆSIn	0.89	1.2
tˆs’	ci	8	ćma	tˆs’ma	0.83	1.2
dˆz’	dzi	X	dźwig	dˆz’vik	0.68	0.7
o+w∼	a_	2	ciąża	ts’ow∼Za	0.63	0.6
c	k	k	kiedy	cjedy	0.50	0.7
dˆz	dz	6	dzwoń	dˆzvon’	0.24	0.2
z’	zi	4	źle	z’le	0.21	0.2
N	N	N	pęk	peNk	0.21	0.1
J	g	g	giełda	Jjewda	0.14	0.1
e+j∼	e_	1	więź	vjej∼s’	0.06	0.1
dˆZ	drz	9	dżem	dˆZem	0.04	0.1

Upodobnienia te dzielimy na[5]:

– **przed spółgłoską dziąsłową** - spółgłoski dziąsłowe (*sz, ż, l, r, dż*) wpływają na spółgłoski przedniojęzykowe-zębowe, przez co zostają zastępowane spółgłoską dziąsłową lub wtórnie udziąsłowioną

– **przed spółgłoską środkowojęzykową** - spółgłoska poprzedzająca spółgłoskę środkowojęzykową (*ś, ć, dź, ź, ń*) staje się jednorodna pod względem artykulacji i miękkości

– **przed zwartymi tylnojęzykowymi** - spółgłoska *n* jest realizowana jako n tylnojęzykowe (w alfabecie IPA jest oznaczana symbolem [η], zaś w alfabecie AGH symbolem [*N*]), przed spółgłoskami zwartymi tylnojęzykowymi (*k*,*g* i ich zmiękczeniami)

Upodobnienia pod względem sposobu artykulacji (w [15] nazywane pod względem zbliżenia narządów) występują gdy zamiast spółgłosek zwartych przed spółgłoskami zwarto-szczelinowymi i szczelinowymi wymawiane są głoski zwarto-szczelinowe(np głoska *t* w słowie *stokrotce*, głoska *d* w słowie *odznaczyć*)

Upodobnienia pod względem dźwięczności występują gdy w wyrazie zachodzi ubezdźwięcznienie lub udźwięcznienie spółgłosek (np. *babka* wymawiamy [bapka], *pośba* wymawiamy *prośba*-[pro4ba]) Najważniejszą zasadą odnoszącą się do tych upodobnień jest: "Grupy złożone ze spółgłosek właściwych wymawia się w całości albo bezdźwięcznie albo dźwięcznie. Decyduje o tym ostatnia spółgłoska w grupie(np. *kredka*-[kretka], *grubszy*-[grupszy])" [6] Zasada ta nie ma zastosowania gdy po spółgłoska bezdźwięcznych występuje dwuznak *rz* lub *w*. Wówczas to *rz,w* ulegają ubezdźwięcznieniu (*trzeba* wymawiamy jako *tszeba*-[tSeba]).

Pliki słowników dla nagrań

Stosując powyższe zasady fonetyki poprawiono błędne transkrypcje fonetyczne wygenerowane przez SARMATĘ. Każdy plik musiał zawierać jednej linii transkrypcje ortograficzną pliku *.wav ze spacjami zastąpionymi znakami "_ " oraz obok transkrypcje fonetyczne tego samego pliku *.wav. Poniżej przykład błędnego i poprawionego kawałka pliku słownika.

Przykład pliku wygenerowanego prze SARMATĘ:

"dzień dobry Xe3dobry

ja chcę zapytać ja x7enzapyta8

żeby być na Dworcu Zachodnim Zebyby8navor7uaxod3im

ale z pętli alezpentlj

tej na Sadybie tejnaadybje

jak mam jechać jakmamjexa8

od ulicy Bonifacego ouli7yo3ifa7ego

...“

Przykład tego samego fragmentu już poprawionego:

“dzień_dobry Xie3dobry

ja_chcę_zapytać ja x7eNapyta8

żeby_być_na_Dworcu_Zachodnim Zebyby8nadvor7uzaxod3im

ale_z_pętli alespeNtli

tej_na_Sadybie tejnasadybje

jak_mam_jechać jakmamjexa8

od_ulicy_Bonifacego oduli7ybo3ifa7ego

...“

3.2. Testy mlf oraz wyniki

3.2.1. Testy baz GMM

Na podstawie tych danych przeprowadzono testy Systemu rozpoznawania mowy SARMATA. Testy zostały przeprowadzone za pomocą wyliczonych modeli GMM oraz porównano wyniki dla modelu GMM przed treningiem mlf oraz dla modeli GMM po treningu dla różnej ilości komponentów (2,3,4,5,6,8,10,12,15,18,20) oraz dwóch metod generowania baz:

1. Bazy wzorców wygenerowano przy użyciu metody, która do wyliczenia modelu używa całego wektora KNN

2. Bazy wzorców wygenerowano przy użyciu metody, która z KNN generuje wektor: [log(en); dct(x)], gdzie

   – log(en) jest to logarytm z energii wektora cech;

   – dct(x) są to cechy KNN poddane transformacji kosinusowej.

   i przy jego pomocy wylicza parametry modelu GMM.

Wyniki testów to procent prawidłowo rozpoznanych fraz ze zbioru testowego. Wyniki baz stworzonych metodą 1 zostały zebrane w tabelach 3.2 i 3.3, metodą drugą zaś w tabelach 3.4 i 3.5. Wykresy 3.1 - 3.14 przedstawiają histogramy baz GMM obu metod, które uzyskały skrajne wyniki (minima, maksima).

Tablica 3.2: Wyniki testów mlf przeprowadzonych dzięki wygenerowanym modelom GMM uzyskanych metodą 1., przed i po treningu mlf wykonanych na zestawach testowych stworzonych z nagrań, na których przeprowadzono trening. Wyróżniono maksima (kolor zielony) i minima (kolor czerwony) .

numer zestawu testowego:	1 [%]	2 [%]	3 [%]	4 [%]	5 [%]	6 [%]	7 [%]	8 [%]	9 [%]	10 [%]
GMM przed treningiem	**30,303**	**60**	46,667	43,750	40	51,724	43,333	56,667	58,065	56,667
GMM_2	78,788	93,333	86,667	84,375	83,333	82,759	73,333	**70**	**93,548**	86,667
GMM_3	78,788	93,333	86,667	84,375	83,333	82,759	73,333	**70**	**93,548**	86,667
GMM_4	78,788	90	80	84,375	76,667	82,759	73,333	**70**	90,323	**96,667**
GMM_5	78,788	93,333	86,667	87,500	80	79,310	**66,667**	**66,667**	90,323	**93,333**
GMM_6	81,818	83,333	80	84,375	86,667	79,310	**76,667**	**76,667**	**93,548**	90
GMM_8	75,758	**70**	**70**	**87,500**	80,000	75,862	70,000	70,000	87,097	76,667
GMM_10	78,788	86,667	76,667	**87,500**	86,667	82,759	**66,667**	**66,667**	87,097	80
GMM_12	84,848	86,667	76,667	87,500	80	86,207	**70**	73,333	90,323	**96,667**
GMM_15	81,818	86,667	80	87,500	83,333	89,655	**73,333**	76,667	**93,548**	86,667
GMM_18	81,818	86,667	83,333	84,375	83,333	86,207	76,667	**66,667**	**87,097**	86,667
GMM_20	81,818	86,667	86,667	81,250	70	86,207	66,667	**63,333**	**90,323**	83,333

Tablica 3.3: Wyniki testów mlf przeprowadzonych dzięki wygenerowanym modelom GMM uzyskanych metodą 1., przed i po treningu mlf wykonanych na zestawach testowych stworzonych z nagrań, na których nie przeprowadzono treningu. Wyróżniono maksima (kolor zielony) i minima (kolor czerwony).

numer zestawu testowego:	1 [%]	2 [%]	3 [%]	4 [%]	5 [%]	6 [%]	7 [%]	8 [%]	9 [%]	10 [%]
GMM przed treningiem	**36,667**	**36,667**	40	51,724	46,667	40	56,667	**60**	46,667	53,333
GMM_2	73,333	66,667	83,333	86,207	76,667	**43,333**	63,333	**93,333**	80	86,667
GMM_3	86,667	80	**90**	79,310	83,333	**43,333**	70	86,667	73,333	83,333
GMM_4	**93,333**	73,333	83,333	89,655	76,667	**46,667**	53,333	93,333	83,333	80
GMM_5	90	80	86,667	86,207	86,667	**46,667**	53,333	90	**93,333**	83,333
GMM_6	90,000	80	86,667	82,759	93,333	**46,667**	56,667	90	**93,333**	83,333
GMM_8	76,667	73,333	70	68,966	86,667	**43,333**	60	**93,333**	76,667	80
GMM_10	90	86,667	93,333	82,759	83,333	**30**	46,667	86,667	**100**	83,333
GMM_12	90	83,333	90	93,103	83,333	**40**	53,333	90	**96,667**	90
GMM_15	86,667	86,667	90	89,655	83,333	**53,333**	56,667	86,667	**93,333**	86,667
GMM_18	80	83,333	90	89,655	83,333	43,333	**36,667**	83,333	**93,333**	**93,333**
GMM_20	76,667	80	90	82,759	83,333	**26,667**	43,333	90	**93,333**	86,667

Tablica 3.4: Wyniki testów mlf przeprowadzonych dzięki wygenerowanym modelom GMM uzyskanych metodą 2., po treningu mlf wykonanych na zestawach testowych stworzonych z nagrań, na których przeprowadzono trening. Wyróżniono maksima (kolor zielony) i minima (kolor czerwony).

numer zestawu testowego:	1 [%]	2 [%]	3 [%]	4 [%]	5 [%]	6 [%]	7 [%]	8 [%]	9 [%]	10 [%]
GMM_2	78,788	93,333	86,667	84,375	83,333	82,759	73,333	**70**	**93,548**	86,667
GMM_3	78,788	93,333	86,667	84,375	83,333	82,759	73,333	**70**	**93,548**	86,667
GMM_4	81,818	**100**	86,667	87,500	90	86,207	**70**	76,667	96,774	93,333
GMM_5	75,758	93,333	80	84,375	90	89,655	**66,667**	73,333	**96,774**	90
GMM_6	81,818	93,333	90	90,625	90	93,103	76,667	**70**	**93,548**	90
GMM_8	81,818	93,333	90	90,625	93,333	89,655	70	**66,667**	**93,548**	86,667
GMM_10	81,818	93,333	93,333	87,500	90	**96,552**	70	**66,667**	93,548	83,333
GMM_12	84,848	93,333	86,667	87,500	90	89,655	**70**	**70**	**96,774**	90
GMM_15	81,818	90	90	**90,625**	83,333	86,207	**60**	70	87,097	90
GMM_18	84,848	90	86,667	90,625	86,667	89,655	**63,333**	66,667	**93,548**	90
GMM_20	81,818	90	93,333	90,625	80	89,655	**63,333**	73,333	**93,548**	86,667

Tablica 3.5: Wyniki testów mlf przeprowadzonych dzięki wygenerowanym modelom GMM uzyskanych metodą 2., po treningu mlf wykonanych na zestawach testowych stworzonych z nagrań, na których nie przeprowadzono treningu. Wyróżniono maksima (kolor zielony) i minima (kolor czerwony).

numer zestawu testowego:	1 [%]	2 [%]	3[%]	4 [%]	5 [%]	6 [%]	7 [%]	8 [%]	9 [%]	10 [%]
GMM_2	73,333	66,667	83,333	86,207	76,667	**43,333**	63,333	**93,333**	80	86,667
GMM_3	90	80	83,333	82,759	80	**40**	63,333	**90**	86,667	80
GMM_4	93,333	80	86,667	86,207	80	**46,667**	70,000	86,667	**93,333**	90
GMM_5	93,333	93,333	83,333	96,552	76,667	**63,333**	70,000	**96,667**	93,333	83,333
GMM_6	93,333	86,667	93,333	82,759	73,333	**50**	66,667	90	**93,333**	83,333
GMM_8	83,333	83,333	**96,667**	93,103	86,667	**56,667**	73,333	**96,667**	**96,667**	90
GMM_10	83,333	86,667	90	93,103	86,667	**50**	63,333	90	**93,333**	90
GMM_12	90	86,667	**96,667**	96,552	90	**53,333**	63,333	86,667	**96,667**	86,667
GMM_15	90	90	86,667	86,207	83,333	**56,667**	60	90	**96,667**	86,667
GMM_18	86,667	93,333	86,667	89,655	86,667	**46,667**	50	86,667	**100**	90
GMM_20	83,333	83,333	90	96,552	86,667	**43,333**	53,333	86,667	**96,667**	93,333

3.2.2. Omówienie wyników

Najlepszą średnią dla baz wygenerowanych metodą 1. otrzymano stosując bazę wzorców GMM z 15 komponentami 3.6 ,3.7, 3.3(zarówno dla zestawów testowych stworzonych z nagrań, na których przeprowadzono trening jak i dla pozostałych zestawów). Najwyższy wynik uzyskano dla zestawu 9 dla bazy GMM z 10 komponentami, z zestawów na których nie przeprowadzono treningów i wynosi on wartość 1. Najwyższy wynik dla zestawów, na których wcześniej przeprowadzano treningi uzyskał zestaw 9 dla bazy GMM z 12 komponentami. Jednak wyraźnie widać, że program lepiej rozpoznawał mowę na zestawach stworzonych z plików, na których trening był przeprowadzony. Zdecydowanie wyższe wówczas miał minima rozpoznawania.

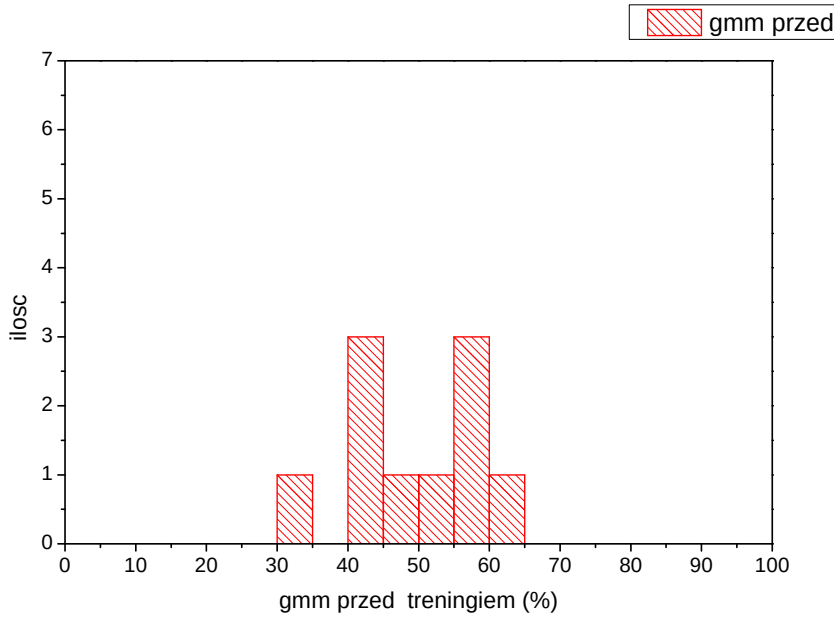
Średnie wyniki dla każdej bazy uzyskanej metodą 2. przedstawiono w tabeli 3.7. Najwyższe średnie zaobserwowano dla baz GMM z 6 komponentami 3.10, oraz dla baz z 8 komponentami 3.13. Można zauważyć że bazy z 8 komponentami dają wyższe średnie dla obu zestawów testowych (tych, na których przeprowadzano i na których nie przeprowadzono treningu). Minima średnio wzrosły o 10% dla zestawów, z których nie korzystano podczas treningu, i nie różnią się zbytnio dla reszty zestawów od wyników baz z metody pierwszej. Średnie dla obu zestawów wzrosty także o średnio 5%.

Tablica 3.6: Uśrednione wyniki oraz wartości maksymalne(max) i minimalne(min) dla baz GMM uzyskanych metodą 1., dla zestawów testowych stworzonych z nagrań, na których przeprowadzono trening(średnia_bg, max_bg, min_bg) oraz dla zestawów testowych stworzonych z nagrań, na których nie przeprowadzono treningu (średnia_nbg, max_nbg, min_nbg). Wyróżniono największe średnie i największe maksima (kolor zielony) i najmniejsze minima-z pominięciem wyników dla bazy GMM wygenerowanej bez treningu (kolor czerwony).

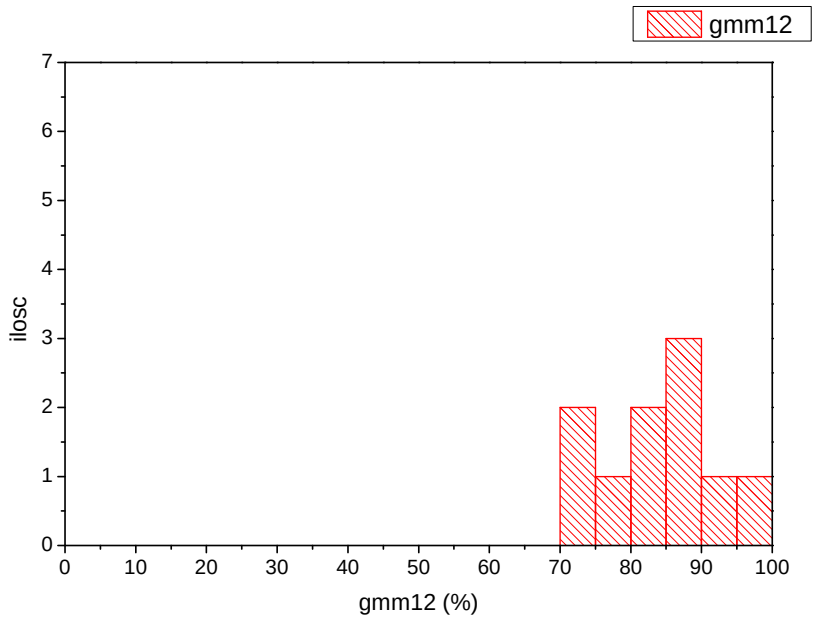
	śr_bg [%]	max_bg [%]	min_bg [%]	śr_nbg [%]	max_nbg [%]	min_nbg [%]
GMM przed treningiem	48,718	60	30,303	46,839	60	36,667
GMM_2	83,280	93,548	70	75,287	93,333	43,333
GMM_3	83,280	93,548	70	77,598	90	43,333
GMM_4	82,291	96,667	70	77,299	93,333	46,667
GMM_5	82,259	93,333	66,667	79,621	93,333	46,667
GMM_6	83,239	93,548	76,667	80,276	93,333	46,667
GMM_8	76,288	87,500	70	72,897	93,333	43,333
GMM_10	79,948	87,500	66,667	78,276	**100**	**30**
GMM_12	83,221	**96,667**	70	80,977	96,667	40
GMM_15	**83,919**	93,548	73,333	**81,299**	93,333	53,333
GMM_18	82,283	87,097	66,667	77,632	93,333	36,667
GMM_20	79,626	90,323	**63,333**	75,276	93,333	26,667

Tablica 3.7: Uśrednione wyniki oraz wartości maksymalne(max) i minimalne(min) dla baz GMM uzyskanych metodą 2., dla zestawów testowych stworzonych z nagrań, na których przeprowadzono trening(średnia_bg, max_bg, min_bg) oraz dla zestawów testowych stworzonych z nagrań, na których nie przeprowadzono treningu (średnia_nbg, max_nbg, min_nbg). Wyróżniono największe średnie, największe maksima (kolor zielony) i najmniejsze minima (kolor czerwony).

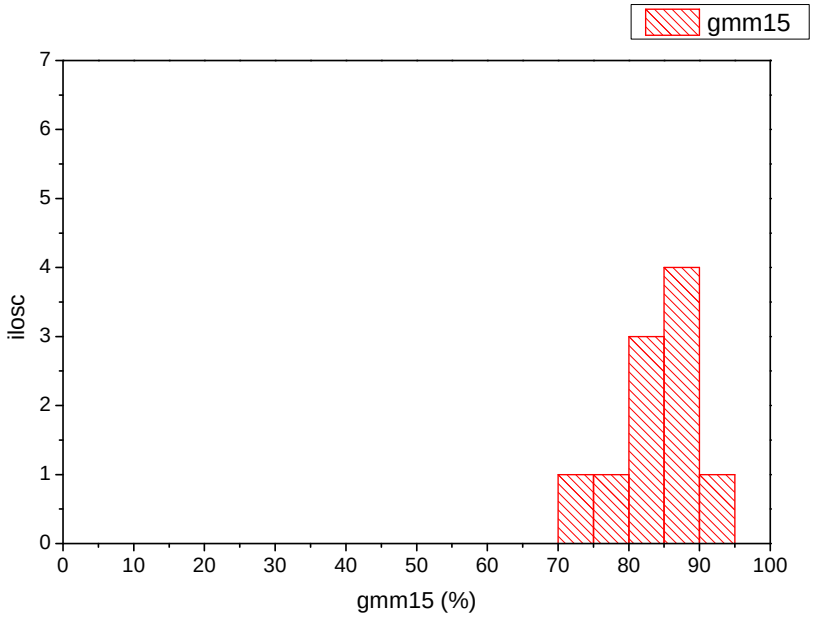
	śr_bg [%]	max_bg [%]	min_bg [%]	śr_nbg [%]	max_nbg [%]	min_nbg [%]
GMM_2	83,280	93,548	70	75,287	93,333	43,333
GMM_3	83,280	93,548	70	77,609	90	**40**
GMM_4	86,897	**100**	70	81,287	93,333	46,667
GMM_5	83,990	96,774	66,667	84,989	96,667	63,333
GMM_6	**86,910**	93,548	70	81,276	93,333	50
GMM_8	85,565	93,548	66,667	**85,644**	96,667	56,667
GMM_10	85,608	96,552	66,667	82,644	93,333	50
GMM_12	85,878	96,774	70	84,655	96,667	53,333
GMM_15	82,908	90,625	60	82,621	96,667	56,667
GMM_18	84,201	93,548	**63,333**	81,632	**100**	46,667
GMM_20	84,231	93,548	63,333	81,322	96,667	43,333



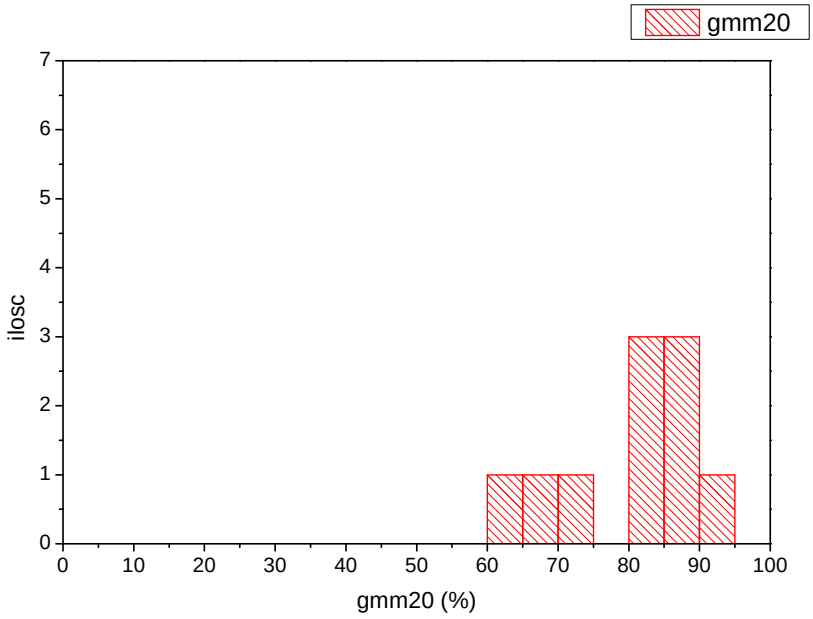
Rysunek 3.1: Histogram wyników dla bazy GMM wygenerowanej przed treningiem, zastosowany na zestawie testowym stworzonym z nagrań, na których przeprowadzano trening mlf.



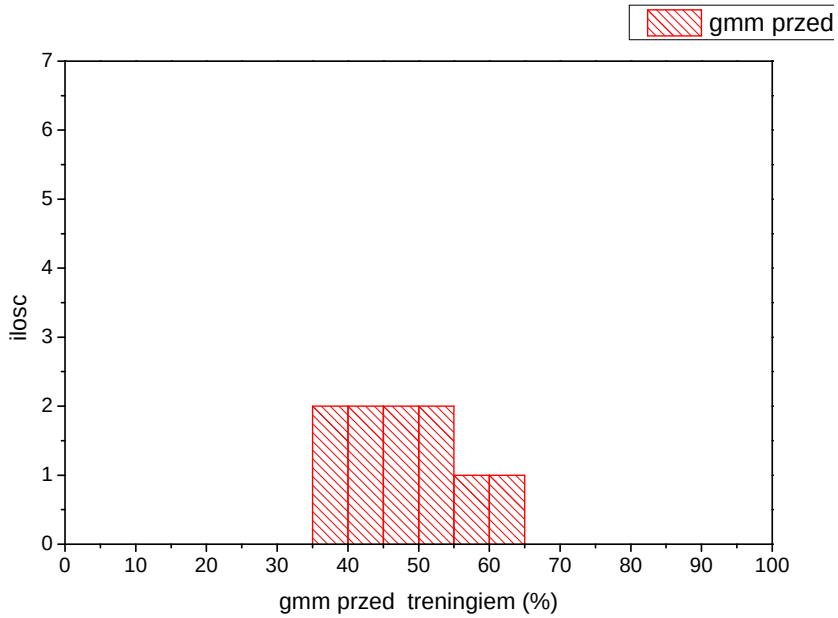
Rysunek 3.2: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 1., z 12 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których przeprowadzano trening mlf. Jest to baza GMM dla której zanotowano maximum dla zestawu nagrań, na których przeprowadzono trening.



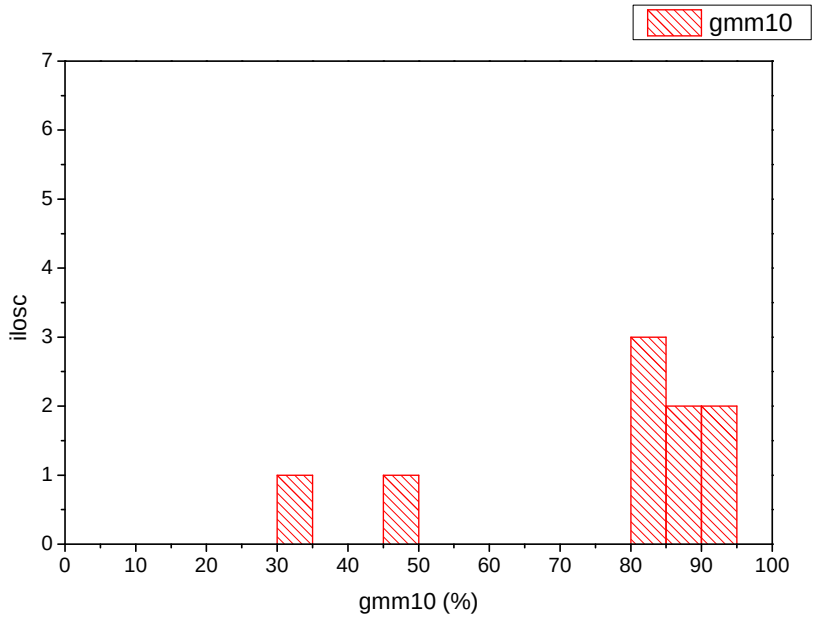
Rysunek 3.3: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 1., z 15 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których przeprowadzano trening mlf.Jest to baza GMM dla której zanotowano najwyższą średnią dla zestawu nagrań, na których przeprowadzono trening



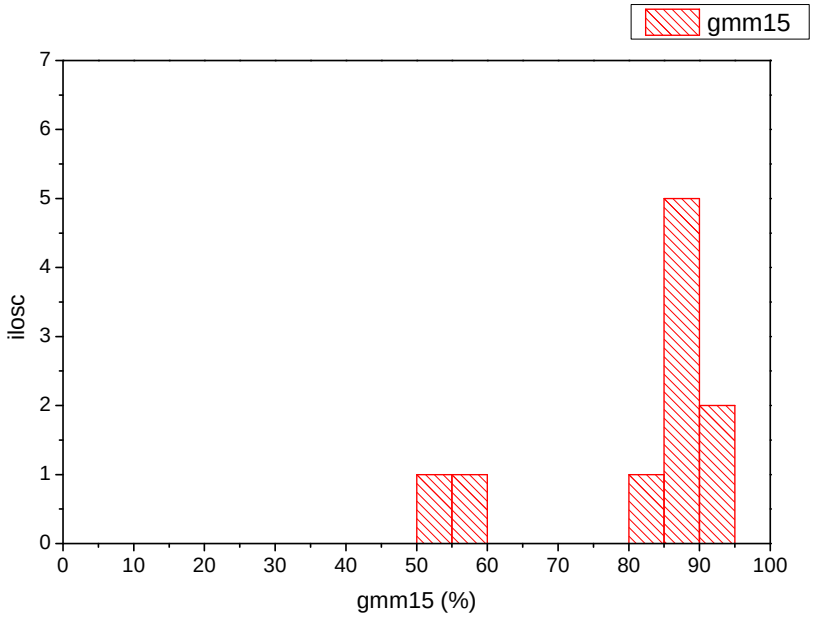
Rysunek 3.4: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 1., z 20 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których przeprowadzano trening mlf. Jest to baza GMM dla której zanotowano minimum dla zestawu nagrań, na których przeprowadzono trening



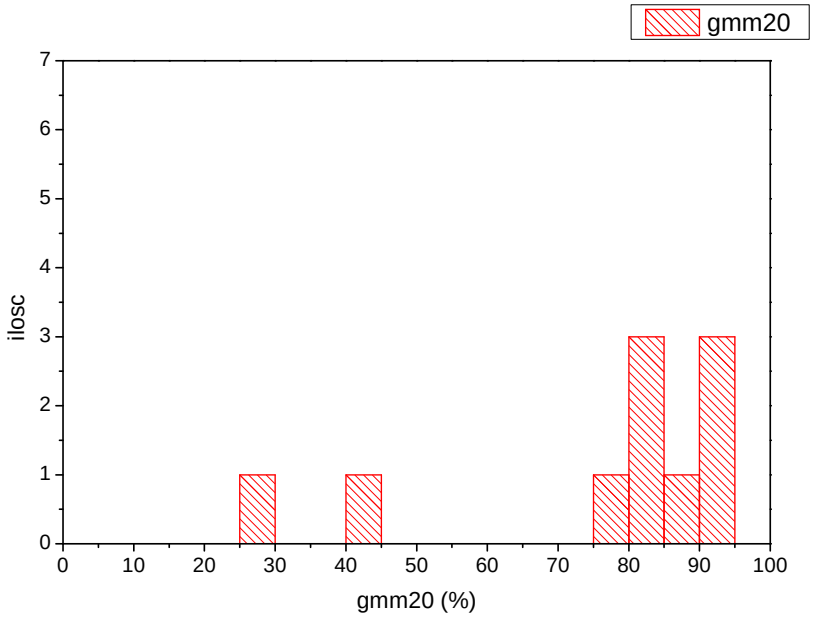
Rysunek 3.5: Histogram wyników dla bazy GMM wygenerowanej przed treningiem, zastosowany na zestawie testowym stworzonym z nagrań, na których nie przeprowadzano treningu.



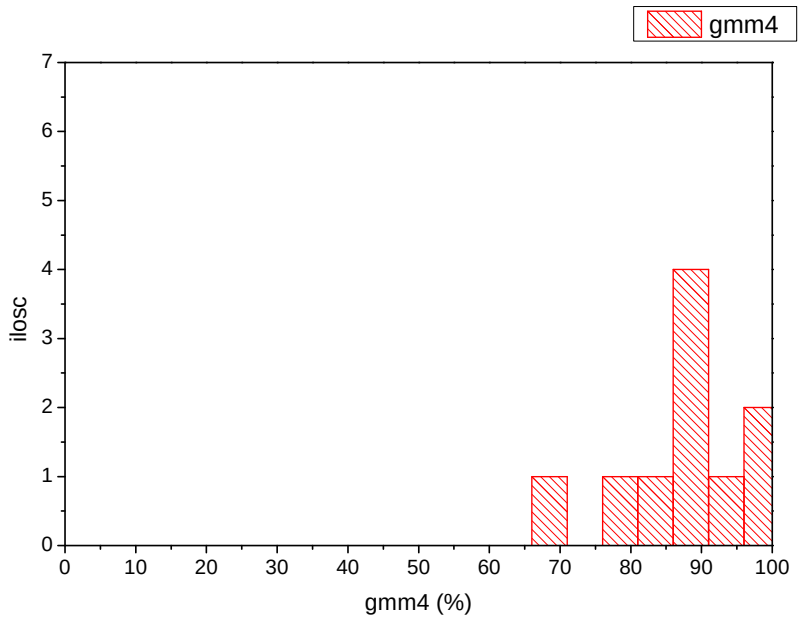
Rysunek 3.6: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 1., z 10 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których nie przeprowadzano treningu. Jest to baza GMM dla której zanotowano maximum dla zestawu nagrań, na których nie przeprowadzono treningu.



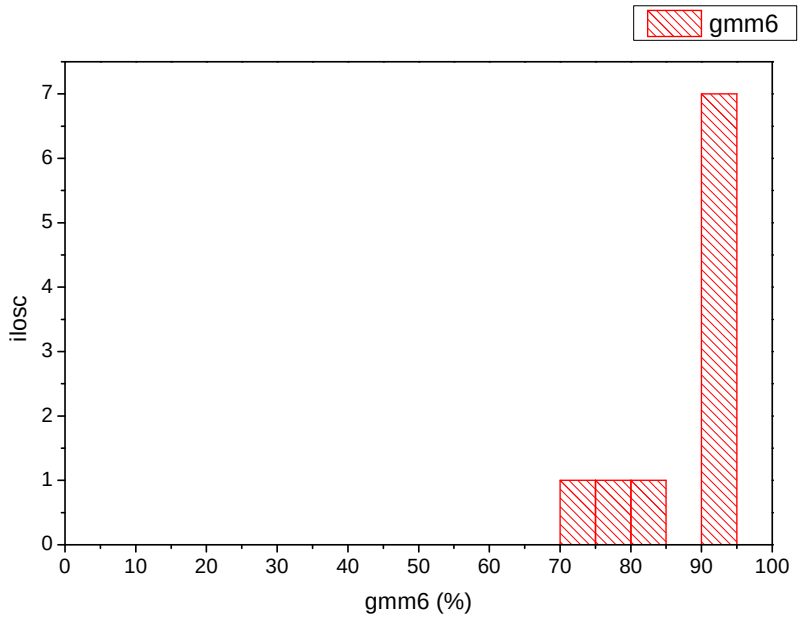
Rysunek 3.7: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 1., z 15 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których nie przeprowadzano treningu. Jest to baza GMM dla której zanotowano najwyższą średnią dla zestawu nagrań, na których nie przeprowadzono treningu.



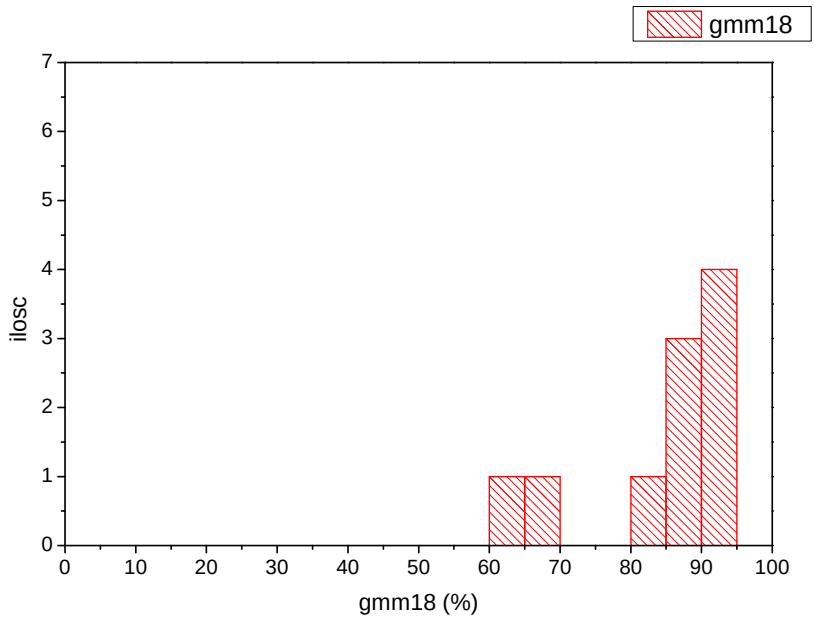
Rysunek 3.8: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 1., z 20 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których nie przeprowadzano treningu. Jest to baza GMM dla której zanotowano minimum dla zestawu nagrań, na których nie przeprowadzono treningu.



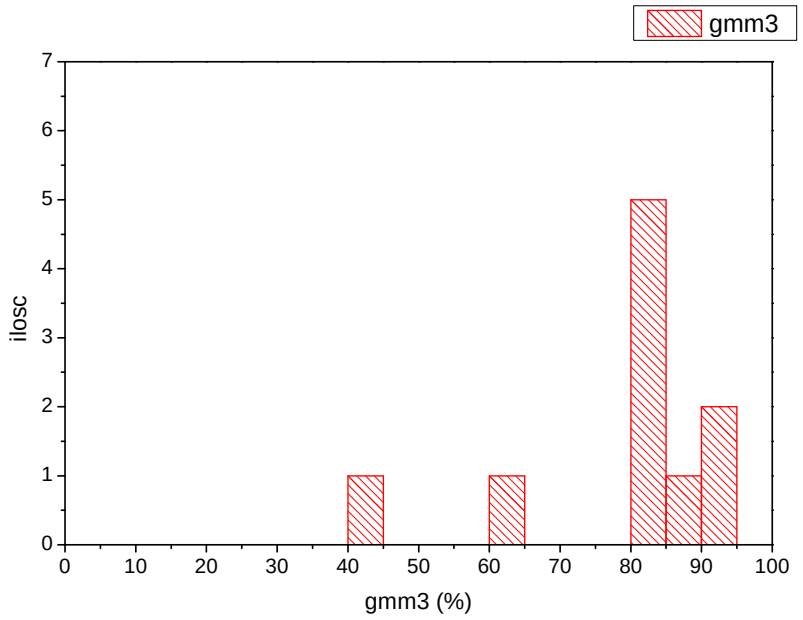
Rysunek 3.9: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 2., z 4 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których przeprowadzono treningu. Jest to baza GMM dla której zanotowano maximum dla zestawu nagrań, na których przeprowadzono trening.



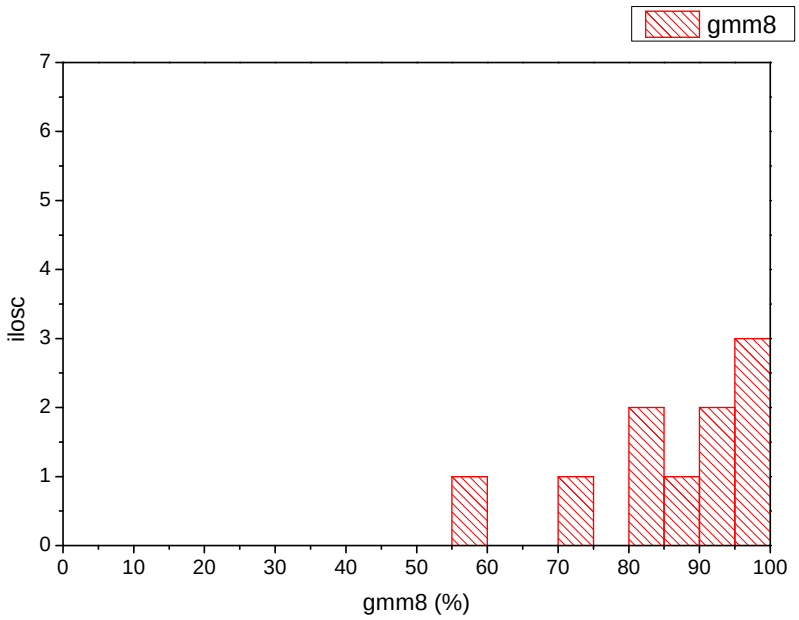
Rysunek 3.10: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 2., z 6 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których przeprowadzono treningu. Jest to baza GMM dla której zanotowano najwyższą średnią dla zestawu nagrań, na których przeprowadzono trening.



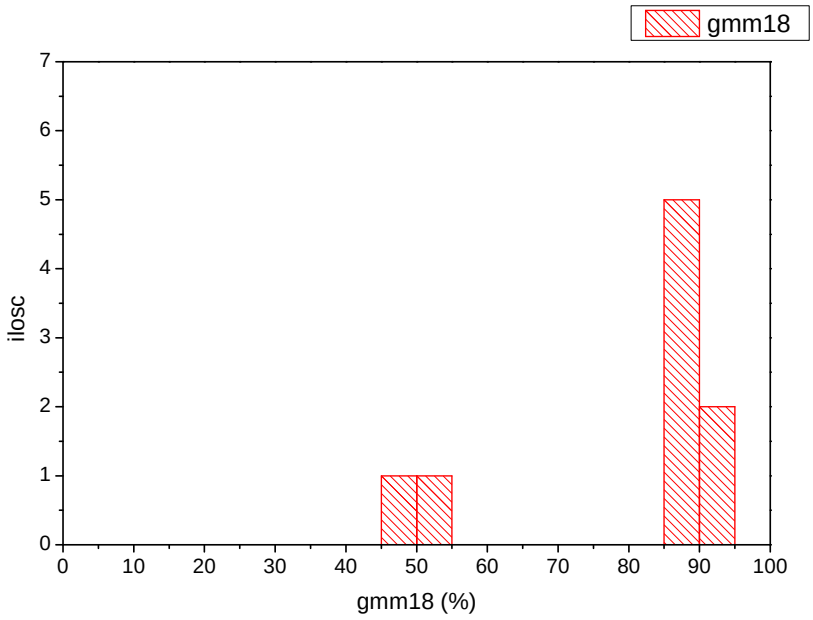
Rysunek 3.11: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 2., z 18 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których przeprowadzono treningu. Jest to baza GMM dla której zanotowano minimum dla zestawu nagrań, na których przeprowadzono trening.



Rysunek 3.12: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 2., z 3 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których nie przeprowadzano treningu. Jest to baza GMM dla której zanotowano minimum dla zestawu nagrań, na których nie przeprowadzono treningu.



Rysunek 3.13: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 2., z 8 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których nie przeprowadzano treningu. Jest to baza GMM dla której zanotowano najwyższą średnią dla zestawu nagrań, na których nie przeprowadzono treningu.



Rysunek 3.14: Histogram wyników dla bazy GMM wygenerowanej dzięki treningowi, uzyskaną metodą 2., z 18 komponentami, zastosowany na zestawie testowym stworzonym z nagrań, na których nie przeprowadzano treningu. Jest to baza GMM dla której zanotowano maximum dla zestawu nagrań, na których nie przeprowadzono treningu.

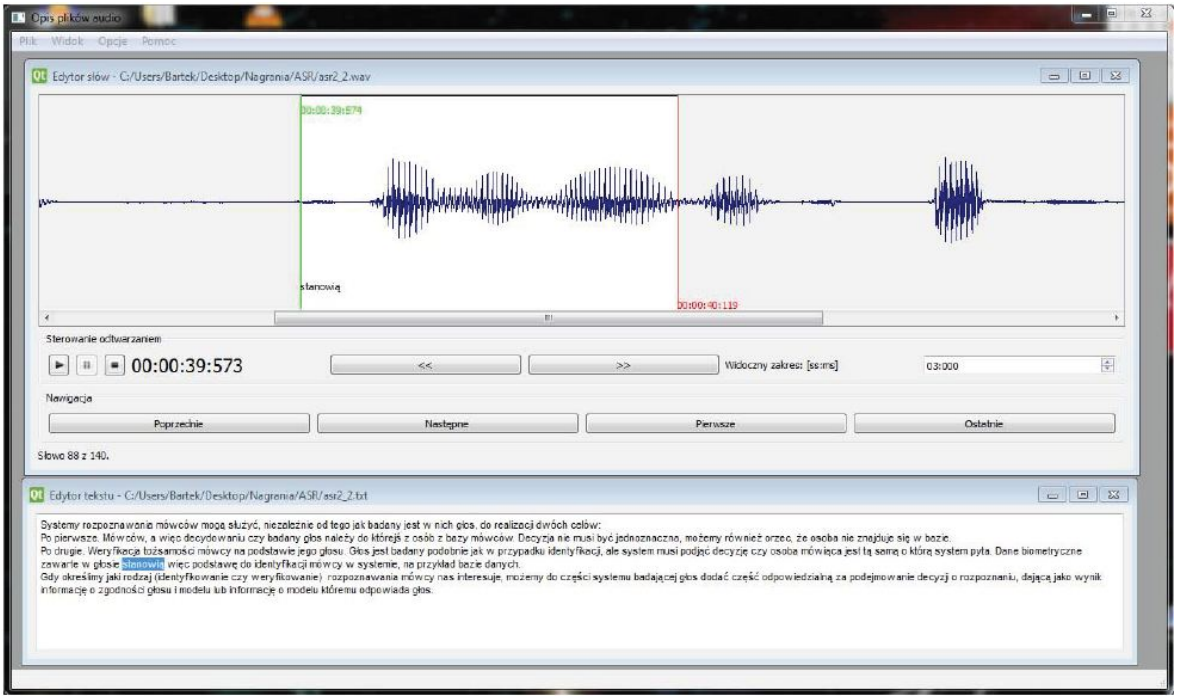
4. Obliczenia sprawności systemu rozpoznawania mowy SARMATA

Program Anotator (rys. 4.1) wykorzystywany był w pracy nad projektem inżynierskim *Rozwinięcie korpusu polskich rozmów telefonicznych LUNA* do tworzenia plików mlf. Tworzenie ich ręcznie było zadaniem żmudnym i bardzo czasochłonnym. Aby usprawnić pracę program został wzbogacony w system SARMATA, który ustawia znaczniki początków i końca słów oraz oznacza fonemy. Osoba tworząca pliki mlf może ustawić program aby rozpoznawał słowa- wówczas oznacza początek i koniec zdania(co ułatwia i skraca czas pracy), lub fonemy, wówczas oznacza początek i koniec słowa. Teraz należało sprawdzić czy SARMATA pracuje odpowiednio i jaka jest jego sprawność w wykrywaniu słów. W tym celu porównano pliki oznaczane ręcznie do plików oznaczanych przez SARMATĘ. Dane pochodziły z korpusu polskich rozmów telefonicznych LUNA oraz z korpusu GlobalPhone.

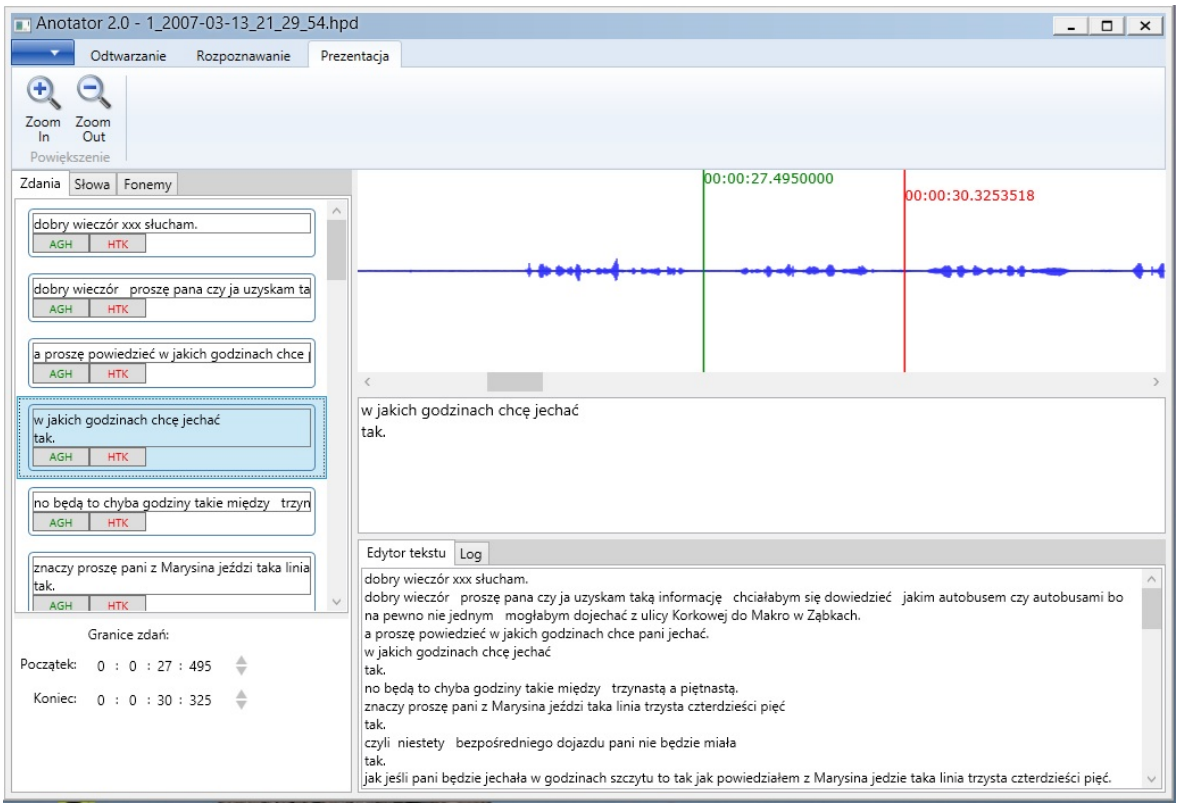
4.1. Porównanie plików mlf oznaczanych ręcznie oraz oznaczanych przez system

Został napisany kod w środowisku MATLAB. Pliki, które są porównywane, muszą znajdować się w dwóch osobnych folderach (folder *"pliki_anotator"* to zbiór plików mlf tworzonych przez Anotator2.0(rys. 4.2), folder *"pliki_recznie"* to zbiór plików tworzonych ręcznie przez człowieka). Kod najpierw odczytuje nazwy plików w obu zbiorach, porównuje czy pliki odnoszą się do tych samych nagrań (poprzez porównanie początków nazw tych plików), następnie, jeśli są to pliki odnoszące się do tego samego nagrania wpisuje ich zawartość do dwóch tabel i je porównuje. Następnie zlicza jak wiele słów:

– zostało podobnie oznaczonych (w kodzie jest możliwość ustawienia parametru błędu dopuszczalnego- o ile milisekund mogą się różnić znaczniki),



Rysunek 4.1: Program Anotator, w którym wykonywano pliki *.mlf [3] w projekcie inżynierskim *Rozwinięcie korpusu polskich rozmów telefonicznych LUNA*



Rysunek 4.2: Program Anotator2.0, wzbogacony o system rozpoznawania mowy

– zostało błędnie oznaczonych,

– nie zostało oznaczonych w ogóle (w plikach oznaczanych przez Anotator2.0 można niestety zauważyć znaczniki czasu "0"- np. *"0 0 się"*- co oznacza że słowo *"się"* zostało nie wykryte).

Pliki mlf muszą być zmodyfikowane w ten sposób, że pomiędzy kolumnami powinien znajdować się znak ";", nie zaś odstęp.

4.1.1. Obliczenie sprawności SARMATY na korpusie polskich rozmów telefonicznych LUNA

Obliczenia i wyniki

Obliczenia przeprowadzono na 22 plikach z korpusu mowy LUNA. Łącznie te nagrania zawierały 9062 słowa. Dobrze oznaczony został czas 269 słów, błędnie oznaczono czas 7808 słów, zaś nie oznaczono 985 słów. Sprawność w tym przypadku wynosiła 0,0297, niepewność pomiaru wynosiła 0,0018. Trzeba jednak zwrócić uwagę, iż LUNA jest bardzo trudnym korpusem mowy, ponieważ zawiera spontaniczne rozmowy człowieka z człowiekiem. Często w nagraniach rozmówcy sobie przerywają lub mówią równocześnie. Zważywszy na to postanowiono przetestować program na korpusie GlobaPhone.

4.1.2. Obliczenie sprawności SARMATY na korpusie GlobalPHONE

Obliczenia i wyniki

Obliczenia przeprowadzano używając tego samego kodu co w przypadku obliczeń na korpusie LUNA. GlobalPhone posiadał załączone pliki mlf, jednakże pliki te zawierają zebrane znaczniki czasu początku i końca słów wszystkich nagrań dla danego mówcy (na około 100 nagrań przypadał jeden plik mlf). Aby porównać pliki anotowane przez SARMATĘ i pliki anotowane ręcznie należało napisać program dzielący te pliki. Napisano go w języku C++, wykorzystując środowisko VisualStudio 2012. Jego zadaniem było podzielenie wczytanego pliku na oddzielne pliki dla każdego nagrania (program w momencie gdy widział kropkę oznaczającą koniec mlf dla kolejnego nagrania rozpoczynał zapisywanie danych do kolejnego pliku). Należało także usunąć znaczniki ciszy z tych plików, gdyż SARMATA nie oznacza dłuższych pauz. W tym celu wykorzystano wyrażenia regularne. Gdy program zauważył linię kończącą się wyrażeniem *"sil"* nie zapisywał jej w pliku wynikowym.

Następnie wykonano anotacje Anotatorem2.0 i wykonano obliczenia za pomocą kodu napisanego w Matlabie.

Obliczenia przeprowadzono na 166 plikach z korpusu mowy GlobalPhone. Nagrania zawierały 2041 słów. Program poprawnie oznaczył czas 21 słów uzyskując sprawność 0,0103 z niepewnością 0,0022. Błędnie oznaczono czas 1829 słów, zaś nie oznaczono 130. W przeciwieństwie więc do zakładanych wyniki są gorsze niż w przypadku korpusu "LUNA".

Złe wyniki ASR załączonego do Anotatora2.0 mogą wynikać z tego, iż jest to wcześniejsze wersja ASR niż testowana w rozdziale 3.

5. VOIP

W ramach pracy magisterskiej wykonano stanowisko VoIP, które posłuży do stworzenia korpusu rozmów telefonicznych.

5.1. Czym jest VoIP

5.1.1. Protokoły internetowe

Protokoły w codziennym życiu oznaczają zbiór zasad i reguł postępowania, jak na przykład protokoły dyplomatyczne. Protokoły internetowe to zbiór zasad i kroków jakie wykonuje komputer podczas komunikacji z innymi komputerami. Stos takich protokołów tworzy model TCP/IP [8].

TCP/IP

Model TCP/IP składa się z czterech warstw ściśle ze sobą powiązanych. Niższe warstwy pobierają dane z warstw wyższych i nie mogą bez nich pracować. Te warstwy to:

- warstwa aplikacji,
- warstwa transportowa,
- warstwa internetu,
- warstwa dostępu do sieci.

Każda z warstw ma własną strukturę i teoretycznie nie zna struktury innych warstw. W rzeczywistości warstwy, jako że współpracują ze sobą, są tak zaprojektowane, aby być kompatybilne, ale wciąż posiadają własną i niezależną strukturę oraz własną terminologię do opisania jej [8][11].

5.1.2. VoIP

VoIP (ang. *Voice Over Internet Protocol*) jest to technologia wykorzystująca protokoły internetowe do przesyłania i odbierania dźwięku. VoIP potocznie nazywany jest telefonią internetową.

Przesyłając dane przez internet odpowiednie oprogramowanie dzieli je na małe pakiety, przesyła każdy pakiet niezależnie, a następnie, w miejscu docelowym, łączy je ponownie w całość. W przypadku pracy VoIP to głos mówcy musi zostać podzielony. Zostaje on zdigitalizowany i zakodowany w jeden lub więcej pakietów przez nadawce, przesyłany przez internet jak każdy inny pakiet danych, a następnie adresat musi go rozkodować[14].

Przez te zabiegi mogą wystąpić problemy będące nie do zaakceptowania dla przeciętnego użytkownika takiej jak: opóźnienie, jitter i utrata pakietów.

Opóźnienie(ang. *delay*)

Opóźnienie to czas jaki pakiet potrzebuje na przebycie drogi z jednego punktu do innego. Aby użytkownik zbytnio nie odczuł opóźnienia i wciąż ono było w granicach tolerancji nie powinno przekraczać 150 ms [14].

Jitter

Jitter to pewien rodzaj opóźnienia- występuje gdy różne pakiety mają różny czas opóźnienia. Gdy jitter jest wysoki lub opóźnienie transmisji jest zbyt wielkie bardzo mocno spada jakość rozmowy.

Utrata pakietów

Utrata pakietów występuje gdy kolejka routera lub limit bufora zostają przekroczone. Może także wystąpić gdy jitter jest zbyt wysoki.

Jednak gdy nasza sieć nie jest przeciążona kodeki dekodujące pakiety VoIP powinny poradzić sobie z powyższymi problemami do tego stopnia, aby użytkownik nie czuł dyskomfortu [14].

5.1.3. Zalety i wady VoIP

Zalety VoIP

Voip pozwala na tanie rozmowy (bezpłatne wewnątrz sieci i zdecydowanie tańsze rozmowy z zagranicą- nie ma roamingu, ponieważ dokładna lokalizacja nie jest ważna do wyko-

nania telefonu- musi mieć jedynie połączenie z internetem). Dzięki VoIP można rozmawiać nie tylko na jednej linii, ale podłączyć do rozmowy wiele linii i tworzyć konferencje. Można także VoIP łączyć z aplikacjami i zarządzać rozmowami lub je zautomatyzować.

Wady VoIP

Główną wadą VoIP jest to, że potrzebny jest dedykowany sprzęt do tego rodzaju usług (np. bramka VoIP). Problemem są także połączenia na numery alarmowe- przez to, że lokalizacja użytkownika, nie jest łatwa do znalezienia w sieci VoIP [14].

5.2. Przygotowanie stanowiska

W ramach projektu Lider uczelnia zakupiła sprzęt VoIP, w skład którego wchodziło:

– router DrayTek,

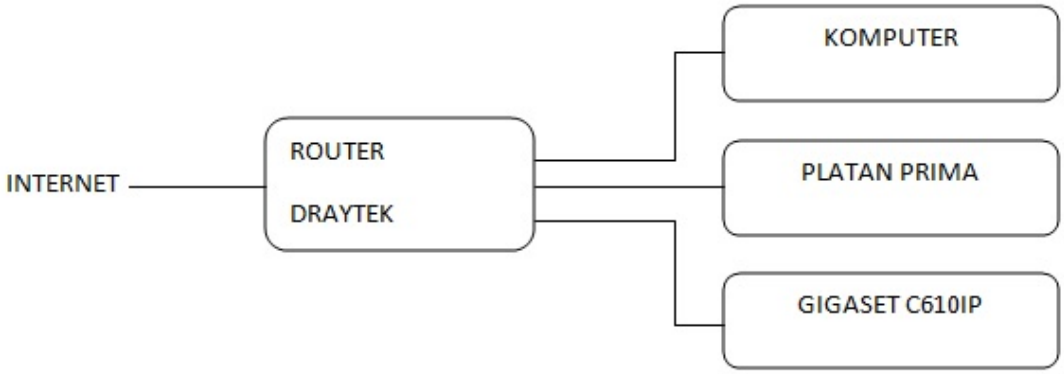
– bramka VoIP Cisco ATA 122,

– telefon Gigaset C610IP,

– centralka VoIP Platan Prima.

Należało przygotować stanowisko, dla którego można napisać aplikację, która będzie dzwoniła do wybranych numerów aby nagrywać rozmowy i w ten sposób stworzyć własny korpus rozmów telefonicznych.

Uczelnia zakupiła dwa numery VoIP, które należało skonfigurować na urządzeniach. Router podłączono do internetu zaś do routera podłączony komputer i reszta urządzeń. Każde z urządzeń było konfigurowane za pomocą przeglądarki internetowej. Należało wprowadzić wszystkie dane kont VoIP i połączyć z siecią tak, aby zostały zaakceptowane. Nie udało się niestety dokonać konfiguracji bramki Cisco ATA 122.



Rysunek 5.1: Schemat połączenia urządzeń VoIP

6. Zakończenie

Systemy rozpoznawania mowy to technologia nadal prężnie się rozwijająca. Wciąż trwają prace nad poprawą działania i sprawności systemów ASR.

W rozdziale trzecim zaprezentowano test baz GMM. Bazy wygenerowano dwoma metodami:

1. Bazy wzorców wygenerowano przy użyciu metody, która do wyliczenia modelu używa całego wektora KNN

2. Bazy wzorców wygenerowano przy użyciu metody, która z KNN generuje wektor: [log(en); dct(x)], gdzie

   – log(en) jest to logarytm z energii wektora cech;

   – dct(x) są to cechy KNN poddane transformacji kosinusowej.

   i przy jego pomocy wylicza parametry modelu GMM.

Z testów wynika, że program zdecydowanie lepiej rozpoznawał mowę na zestawach stworzonych z nagrań, na których przeprowadzono trening. Dla baz GMM stworzonych metodą 2 minima średnio wzrosły o 10% dla zestawów, z których nie korzystano podczas treningu, i nie różnią się zbytnio dla reszty zestawów od wyników baz z metody pierwszej. Średnie dla obu zestawów wzrosty także o średnio 5%.

W rozdziale czwartym przedstawiono test programu Anotator2.0. Niestety ASR załączony do programu nie spełnia oczekiwań. System oznacza słowa podobnie do oznaczenia ręcznego jedynie w 2,9 % (wyniki dla korpusu LUNA) oraz w 1% (wyniki dla korpusu GlobalPHONE). Złe wyniki ASR załączonego do Anotatora2.0 mogą wynika z tego, że jest to wczesna wersja ASR.

W rozdziale piątym zaprezentowano technologię VoIP, jej wady i zalety oraz omówiono wykonanie stanowiska VoIP, które posłuży do nagrywania rozmów telefonicznych i dzięki temu będzie można stworzyć korpus rozmów telefonicznych.

Bibliografia

[1] J. Gałka M. Ziółko B. Ziółko, D. Skurzok. Speech modelling based on phone statistics. *2010 Fifth International Multi-conference on Computing in the Global Information Technology, Valence, Spain*, 2010.

[2] M. Ziółko B. Ziółko. *Przetwarzanie Mowy*. Wydawnictwo AGH, 2011.

[3] T. Jadczyk B. Ziółko, B. Miga. Semisupervised production of speech corpora using existing recordings. *Proceedings of International Seminar on Speech Production (ISSP'11), Montreal*, 2011.

[4] C. Basztura. *Rozmawiać z komputerem*. Format, 1992.

[5] J. Tambor D. Ostaszewska. *Fonetyka i fonologia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN SA, 2000.

[6] B. Dunaj. Zasady poprawnej wymowy polskiej. *Język Polski LXXXVI*, zeszyt 3:162–172, 2006.

[7] J. Gałka. *Optymalizacja parametryzacji sygnału w aspekcie rozpoznawania mowy polskiej*. PhD thesis, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, 2008.

[8] C. Hunt. *TCP/IP Network Administration, Third Edition*. O'Reilly Media, 2002.

[9] B. Ziółko T. Jadczyk D. Skurzok M. Mąsior M. Ziółko, J. Gałka. Automatic speech recognition system dedicated for polish. *Show and tell session, Interspeech 2011, Florencja.*

[10] M. Marciniak. *Anotowany korpus dialogów telefonicznych*. Exit, 2010.

[11] P. Miller. *TCP/IP: the ultimate protocol guide*. BrownWalker Press, 2009.

[12] T. Schultz N. Thang Vu, F. Kraus, editor. *Multilingual a-stabil: a new confidence score for multilingual unsupervised training*. IEEE Workshop on Spoken Language Technology, SLT, 2010.

[13] D. Reynolds. *Encyclopedia of Biometrics*. Springer Publishing Company, Incorporated, 2009.

[14] S. Bhatnagar S. Ganguly. *VoIP: Wireless, P2P and New Enterprice Voice over IP*. Chicester: John Wiley & Sons, 2008.

[15] K. Szklanny. *Optymalizacja funkcji kosztu w korpusowej syntezie mowy polskiej*. PhD thesis, Polsko-Japońska Wyższa Szkoła Technik Komputerowych, 2009.

[16] S. Young, G. Evermann, M. Gales, Th. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *HTK Book*. Cambridge University Engineering Department, UK, 2005.