AKADEMIA GÓRNICZO-HUTNICZA

im. Stanisława Staszica w Krakowie

WYDZIAŁ INŻYNIERII MECHANICZNEJ I ROBOTYKI

Praca dyplomowa
inżynierska

Bartosz Stoliński
Imię i nazwisko

Inżynieria akustyczna
Kierunek studiów

System detekcji nagrań
Temat pracy dyplomowej

Dr inż. Bartosz Ziółko
Promotor pracy

……..
Ocena

Kraków, rok 2013/2014

Kraków, dnia .....

Imię i nazwisko: Bartosz Stoliński

Nr albumu: 241193

Kierunek studiów: **Inżynieria Akustyczna**

Profil dyplomowania: -

OŚWIADCZENIE

Świadomy odpowiedzialności karnej za poświadczanie nieprawdy oświadczam, że niniejszą inżynierską pracę dyplomową wykonałem osobiście i samodzielnie oraz nie korzystałem ze źródeł innych niż wymienione w pracy.

Jednocześnie oświadczam, że dokumentacja praca 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 uzyskałem w sposób niedozwolony. Wersja dokumentacji dołączona przeze mnie na nośniku elektronicznym jest w pełni zgodna z wydrukiem przedstawionym do recenzji.

Zaświadczam także, że niniejsza inżynierska praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów wyższej uczelni lub tytułów zawodowych.

……..

podpis dyplomanta

Kraków, ……..

Imię i nazwisko: Bartosz Stoliński

Adres korespondencyjny: os. Władysława Jagiełły 5/6, 32-800 Brzesko

Temat pracy dyplomowej inżynierskiej: System detekcji nagrań

Rok ukończenia: 2014

Nr albumu: 241193

Kierunek studiów: Inżynieria Akustyczna

Profil dyplomowania: -

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 inżynierskiej pracy dyplomowej, w zakresie publicznego udostępniania i rozpowszechniania w wersji drukowanej i elektronicznej[1].

Publikacja ta może nastąpić po ewentualnym zgłoszeniu do ochrony prawnej wynalazków, wzorów użytkowych, wzorów przemysłowych będących wynikiem pracy inżynierskiej[2].

Kraków, 23.01.2014 ……..
data *podpis dyplomanta*

[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."

[2] Ustawa z dnia 30 czerwca 2000r. – Prawo własności przemysłowej (Dz.U. z 2003r. Nr 119, poz. 1117 z późniejszymi zmianami) a także rozporządzenie Prezesa Rady Ministrów z dnia 17 września 2001r. w sprawie dokonywania i rozpatrywania zgłoszeń wynalazków i wzorów użytkowych (Dz.U. nr 102 poz. 1119 oraz z 2005r. Nr 109, poz. 910).

Kraków, dnia .....

AKADEMIA GÓRNICZO-HUTNICZA
WYDZIAŁ INŻYNIERII MECHANICZNEJ I ROBOTYKI

TEMATYKA PRACY DYPLOMOWEJ INŻYNIERSKIEJ
dla studenta IV roku studiów stacjonarnych

Bartosz Stoliński
imię i nazwisko studenta

TEMAT PRACY DYPLOMOWEJ INŻYNIERSKIEJ:

System detekcji nagrań

Promotor pracy: dr inż. Bartosz Ziółko

Recenzent pracy: dr hab. inż. prof. AGH Piotr Kleczkowski *Podpis dziekana:*

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

Kraków, ...… ……..
data *podpis dyplomanta*

TERMIN ODDANIA DO DZIEKANATU: ..... 2014 r.

.....
podpis promotora

Akademia Górniczo-Hutnicza im. Stanisława Staszica Kraków, 23.01.2014

Wydział Inżynierii Mechanicznej i Robotyki

Kierunek: Inżynieria Akustyczna

Bartosz Stoliński

Praca dyplomowa inżynierska

System detekcji nagrań

Opiekun: dr inż. Bartosz Ziółko

STRESZCZENIE

W pracy omówiono techniki rozpoznawania nagrań i sprawdzono wybrane spośród nich pod kątem ich przydatności w systemie detekcji poczty głosowej. Spośród omówionych technik wybrano: ekstrakcję cech kanału transmisyjnego za pomocą PFCC, detekcję nagrań z wykorzystaniem trenowanego klasyfikatora formantów, detekcję różnic w kanale transmisyjnym oraz rozpoznawanie mowy. W większości testów udało się uzyskać wiarygodne wyniki i potwierdzić skuteczność testowanych technik w kontekście detekcji poczty głosowej. W podsumowaniu pracy zawarto również wskazówki odnośnie możliwości implementacyjnych każdej z technik, doboru parametrów oraz możliwych dalszych kierunków badań.

AGH University of Science and Technology Kraków, the 23.01.2014

Faculty of Mechanical Engineering and Robotics

Field of Study: Acoustic Engineering

Bartosz Stoliński

Engineer Diploma Thesis

Recorded speech detection system

Supervisor: dr inż. Bartosz Ziółko

SUMMARY

The methods of detecting a recorded speech were analysed and tested according to their applicability in the field of voicemail detection in this thesis. Methods chosen for testing were: transmission channel characteristics extraction with PFCC, recorded speech detection with trained pattern classifier, differences in transmission channels and speech recognition. Most of the tests gave results credible enough to confirm methods' usefulness in the field of voicemail detection. Suggestions of implementation possibilities and parameters of each method and possible trends of further studies were also included.

SPIS TREŚCI

1. WSTĘP

1.1. Wprowadzenie.

W dynamicznie rozwijającym się rynku usług telekomunikacyjnych i postępie w automatycznej obsłudze klienta niezbędną cechą każdego systemu staje się umiejętność zweryfikowania autentyczności jednostki odbierającej połączenie. W automatycznym telemarketingu, w którym system sam dzwoni do klienta z ofertą, możliwość sprawdzenia przez system, czy dodzwonił się do prawdziwej osoby jest kluczową kwestią w ograniczeniu kosztów i czasu traconego na niepotrzebne połączenia z pocztą głosową lub przedstawicielem innej firmy. Pomysł na zajęcie się tą kwestią autorowi podsunął jeden z pracowników krakowskiej firmy telemarketingowej, w której informatycy nie mogli znaleźć skutecznego rozwiązania problemu.

1.2. Cel pracy.

Celem niniejszej pracy jest analiza przydatności możliwych technik detekcji nagrań w kontekście wykrywania poczty głosowej oraz ich implementacja i weryfikacja uzyskanych wyników.

1.3. Opis rozdziałów.

W niniejszej pracy omówione i przetestowane zostały techniki detekcji nagrań z wyszczególnieniem wykrywania nagrań odtwarzanych w trakcie połączenia telefonicznego.

Rozdział 2. zawiera część analityczną, w której omówiono problem detekcji nagrań w kontekście poczty głosowej, możliwe sposoby rozwiązania problemu oraz opis narzędzi wykorzystanych do przeprowadzenia badań. Rozdział 3. zawiera szczegółowy opis każdej z wybranych metod detekcji nagrań oraz zastosowanych algorytmów. W rozdziale 4. zestawiono i omówiono uzyskane wyniki. Rozdziały 5 i 6 to kolejno: podsumowanie pracy oraz wnioski i bibliografia.

2. ANALIZA ZAGADNIENIA

2.1. Problem detekcji nagrań w kontekście poczty głosowej.

Detekcję poczty głosowej można potraktować jako szczególny przypadek detekcji nagrań. Aby znaleźć rozwiązanie dokonano zatem analizy metod rozpoznawania nagrań, oceniono ich przydatność w omawianym kontekście i wybrano najbardziej obiecujące spośród nich. Głównym wymogiem systemu jest minimalizacja czasu weryfikacji, natomiast skuteczność algorytmów związanych z analizą mowy rośnie wraz z ilością informacji, czyli czasem trwania wypowiedzi. Konieczne zatem staje się zbalansowanie skuteczności weryfikacji z czasem reakcji systemu. Zbyt długi czas analizy może skutkować zakończeniem połączenia przez odbiorcę, spowodowanym brakiem reakcji, zbyt krótki natomiast może błędnie zweryfikować i odrzucić potencjalnego klienta. Umożliwienie użytkownikowi zmiany progu rozpoznania będzie dodatkowym atutem systemu, zwiększającym jego elastyczność. W dalszej części rozdziału omówiono możliwe techniki dystynkcji nagrań od rzeczywistej mowy i ich potencjał w kontekście detekcji poczty głosowej.

2.2. Możliwe rozwiązania.

2.2.1. Analiza metod.

Mammone i Sharma zaproponowali kilka różnych sposobów wykrywania nagrań [3]. W następnych rozdziałach opisano wybrane techniki, przetestowano i omówiono uzyskane wyniki oraz oceniono przydatność każdej z nich w kontekście wykrywania poczty głosowej.

Jedną z proponowanych metod jest badanie czasowych charakterystyk głosu [3]. Dokładne powtórzenie danej wypowiedzi jest niemożliwe, mimo iż dla ludzkiego ucha dwie wypowiedzi mogą brzmieć identycznie. Komputer jednak jest w stanie wykryć różnice między takimi wypowiedziami. W zastosowanym rozwiązaniu system prosi użytkownika o podanie hasła, a następnie porównuje wypowiedź z nagraniami z bazy danego użytkownika. Gdy dwie wypowiedzi są do siebie zbyt podobne system charakteryzuje wypowiedź jako nagranie. Zastosowane w rozwiązaniu techniki porównania dwóch wypowiedzi to min. przejścia przez zero, obwiednia sygnału i czas trwania właściwej wypowiedzi. Rozwiązanie jest skuteczne w identyfikowaniu nagrań

i proste w implementacji, jednak w kontekście poczty głosowej, jest niemożliwe do zastosowania, gdyż opiera się na zbieraniu informacji na temat użytkownika.

Inną proponowaną techniką jest detekcja różnic w kanale transmisyjnym [3]. Opiera się ona na zbieraniu informacji na temat kanału transmisyjnego zawartych w pozyskanej próbce dźwięku i porównaniu z inną próbką dźwięku uzyskaną w czasie trwania jednego połączenia. Jest to skuteczna metoda w sytuacji gdy użytkownik wypowiada się więcej niż jeden raz, jednak jej zastosowanie w detekcji automatycznych sekretarek może okazać się niemożliwe. Problemem jest ilość informacji potrzebnych do uzyskania wiarygodnej estymacji kanału, czyli czas trwania wypowiedzi. Ideą systemu detekcji poczty głosowej jest minimalizacja kosztów, poprzez redukcję czasu trwania niechcianych połączeń. Istotny jest zatem jak najkrótszy czas rozpoznania. Zastosowanie techniki detekcji kanału do wykrywania automatycznych sekretarek może być możliwe w przypadku opracowania dokładniejszych algorytmów estymacji kanału.

Przy połączeniu z pocztą głosową, nagranie nie jest odtwarzane natychmiastowo. Między uzyskaniem połączenia, a odtworzeniem nagrania zawsze jest różniący się czasem trwania przedział. Gdyby udało się stworzyć wystarczająco skuteczny algorytm do estymacji kanału na krótkim odcinku czasu, możliwe byłoby porównanie cech kanału zebranych przed wypowiedzią i w trakcie jej trwania.

Kolejnym proponowanym rozwiązaniem jest wykorzystanie trenowanego klasyfikatora formantów. Klasyfikator wymaga treningu bazą danych zawierającą nagrania oryginalne i odtworzone. Stworzenie odpowiednio dużej bazy danych jest proste w realizacji, istnieje również dowolność w wyborze klasyfikatora. Proponowane klasyfikatory to np. GMM (ang. *Gaussian Mixture Model*) [12], HTK (ang. *Hidden Markov Model Toolkit*) [9], SVM (ang. *Support Vector Machines*) [11] i NTN (ang. *Neural Tree Networks*) [11]. Ponieważ bazę nagrań można stworzyć na zewnętrznej grupie, technika może być wykorzystana do wykrywania poczty głosowej. Problemem może być dobranie odpowiednio szybkiego klasyfikatora.

Innym możliwym rozwiązaniem jest zastosowanie cyfrowego znaku wodnego [3]. Opiera się ona na oznaczaniu nagrań sekwencją impulsów o ustalonej częstotliwości. W większości urządzeń posiadających mikrofon i głośnik (np. telefon komórkowy/stacjonarny, zestaw słuchawkowy z mikrofonem) występuje sprzężenie zwrotne. W omawianym rozwiązaniu właściwość ta jest wykorzystana, w celu

przechwycenia nagrania z sygnałem wysłanym do użytkownika. Dzięki tej technice odebrane nagranie zyskuje unikalne oznaczenie. Jeżeli w trakcie innego połączenia zostanie w nagraniu wykryty znak wodny znajdujący się w bazie danych system odmawia dostępu. Jest to skuteczne rozwiązanie w systemach weryfikacji tożsamości jednak ponieważ opiera się na wewnętrznej bazie danych nie jest przydatne w systemie detekcji poczty głosowej. Możliwym byłoby jednak wykorzystanie elementów opisanej techniki. Przyjmując, że poczta głosowa odtwarza nagranie z nośnika pamięci, nie jest możliwym sprzężenie zwrotne między odbiornikiem i nadajnikiem. Nadając zatem wygenerowany sygnał można spodziewać się braku odpowiedzi w przypadku połączenia się z pocztą głosową. Rozwiązanie to wymagałoby jednak skutecznej metody przechwytywania sygnału sprzężenia zwrotnego. Ze względu na złożoność zagadnienia i trudność w realizacji metody testowania, metoda ta nie została zbadana w poniższej pracy.

Odmiennym podejściem do problemu detekcji nagrań jest rozpoznawanie mowy. Rozwiązanie to jest nieskuteczne przy porównaniu dwóch plików dźwiękowych, lecz w przypadku połączeń telefonicznych można wykorzystać je wykorzystać opierając się o statystykę. Słowa wypowiedziane przez osobę odbierającą połączenie są unikalne w stosunku do słów wypowiadanych przy odtworzeniu nagrania. Możliwym rozwiązaniem jest sprawdzanie np. dwóch pierwszych uzyskanych słów i na ich podstawie ustalenie czy połączono się z żywym użytkownikiem, czy pocztą głosową.

Możliwe są również rozwiązania, takie jak badanie długości przerw między słowami (prosta i obiecująca metoda, gdyż w przypadku odebrania połączenia następuje dłuższa cisza po pierwszych słowach), czy też sprawdzanie ilości wypowiedzianych słów (przy odebraniu połączenia najczęściej wypowiadane są jedno lub dwa słowa, większą ilość można uznać za odtworzone nagranie).

2.2.2. Wykorzystane narzędzia.

2.2.2.1. MATLAB

Środowisko MATLAB zostało wykorzystane do implementacji rozwiązań opartych na badaniu właściwości czasowych i widmowych sygnałów. O wyborze zadecydowała wygoda i krótki czas potrzebny na implementację, możliwe dzięki wysokiemu poziomowi abstrakcji składni oprogramowania oraz bardzo dużej bazie wbudowanych funkcji.

Dla przyspieszenia pracy nad kodem wykorzystano rozszerzenie Voicebox, zawierające zbiór funkcji pomocnych w analizie mowy.

2.2.2.2. Simulink

Do symulacji kanału transmisyjnego wykorzystano program Simulink, będący częścią pakietu MATLAB. O wyborze zadecydowała możliwość bezpośredniej wymiany danych między programami.

2.2.2.3. Sarmata

Do implementacji techniki opartej na identyfikacji wypowiadanych słów wybrano stworzony przez zespół cyfrowego przetwarzania sygnałów AGH program Sarmata. O wyborze zadecydowała dostępność oprogramowania oraz zawarty w nim duży korpus mowy polskiej.

2.2.2.4. Phoner

Do wykonania nagrań oraz statystyki słów wypowiadanych przy odbieraniu połączeń wykorzystano program Phoner. Program wybrano dzięki łatwej dostępności oraz wbudowanej opcji nagrywania połączeń.

2.2.2.5. Nagrania

Do symulacji i testów wykorzystano dwa zestawy nagrań. Pierwszy zestaw to nagrania wykonany programem Phoner podczas połączeń telefonicznych. Zestaw składa się ze 123 nagrań o różnym czasie trwania, został podzielony na dwie części, w zależności czy połączenie odebrała żywa osoba, czy zgłosił się automat.

Drugi zestaw to część pakietu nagrań, wykonanych przez autora na potrzeby pracy, podczas połączenia za pomocą telefonii cyfrowej, udostępnionego na potrzeby pracy przez zespół przetwarzania sygnałów cyfrowych AGH, składająca się ze 162 nagrań, zawierających tę samą wypowiedź. Nagrania zostały wykonane na dwa sposoby: mikrofonem zewnętrznym oraz po stronie serwera. Połowa nagrań zawiera czysty sygnał mowy (mikrofon), a połowa sygnał zakłócony kanałem transmisyjnym (serwer).

3. OPIS PRZYJĘTYCH ROZWIĄZAŃ

Spośród technik opisanych w rozdziale drugim wybrano, opisano i zbadano najbardziej obiecujące i będące możliwymi do realizacji przy pomocy dostępnych narzędzi oraz poziomu posiadanej wiedzy i umiejętności.

3.1. Detekcja cech kanału przy pomocy predykcji liniowej

Predykcja liniowa jest potężnym narzędziem dającym szerokie możliwości analizy i przetwarzania sygnału mowy. Za jej pomocą można przeprowadzać analizę widmową sygnału, uzyskać obwiednię sygnału, formanty, a także pozwala efektywnie przesyłać sygnał, jest również wykorzystywana w przetwarzaniu mowy [5], [10], [13], [14], [15]. Dzięki przeprowadzanej za pomocą predykcji liniowej analizie cepstralnej można skutecznie estymować sygnał ekscytacji kanału głosowego, bez wpływu traktu głosowego [6] lub zadziałać odwrotnie i dokonać ekstrakcji cech kanału głosowego. W dalszej części rozdziału posłużono się analogią wpływu kanału głosowego na wzbudzenie krtaniowe do wpływu kanału transmisyjnego na transmitowany sygnał.

Sygnał mowy $x(t)$ można przedstawić w postaci splotu sygnału wzbudzającego (tonu krtaniowego) $g(t)$ z układem o zmiennej odpowiedzi impulsowej (tor głosowy) $h(t)$ [4]. Splot w dziedzinie czasu można przedstawić jako iloczyn widm sygnałów:

$$X(f) = G(f)H(f) \tag{3.1}$$

Analogicznie do (3.1) podczas transmisji sygnał mowy o widmie $S(f)$ jest zakłócany przez kanał transmisyjny o widmie $X(f)$. Widmo sygnału zakłóconego $\hat{S}(f)$ dane jest wzorem:

$$\hat{S}(f) = S(f)X(f) \tag{3.2}$$

Korzystając z własności logarytmów oraz splotu można zapisać powyższe równanie w postaci:

$$\mathcal{F}^{-1}\ln\left|\hat{S}(f)\right| = \mathcal{F}^{-1}\ln|S(f)| + \mathcal{F}^{-1}\ln|X(f)| \tag{3.3}$$

Zgodnie z definicją cepstrum [7] powyższe równanie można przedstawić w dziedzinie cepstralnej jako:

$$\hat{c}(n) = c(n) + x(n) \tag{3.4}$$

Niezmienne w czasie zakłócenie (estymację kanału transmisyjnego) $X(f)$ można wyeliminować poprzez uśrednianie sygnału w dziedzinie cepstralnej i odjęcie uzyskanej wartości średniej. Przy założeniu, że średnia wartość składnika równania cepstralnego odpowiadającego sygnałowi mowy $c(n)$ dąży do zera, estymacja kanału transmisyjnego jest równoważna średniej wartości $\hat{c}(n)$.

W praktyce jednak średnia cepstralna może zawierać nie tylko informacje o kanale, lecz również elementy widma mowy [1], [2], [3], [6]. Technika filtracji biegunowej (ang. *Pole Filtering*) wprowadza dodatkowe elementy do algorytmu estymacji kanału metodą średniej cepstralnej, oparte na manipulacji biegunami filtra predykcyjnego, mające na celu ograniczenie części widma sygnału mowy zawartego w średniej cepstralnej.

Cepstrum można przedstawić jako ważoną kombinację współczynników (lub biegunów) filtra uzyskanego metodą predykcji liniowej (LP). Wynikiem LP jest filtr $A(z)$ mający wszystkie pierwiastki wewnątrz koła jednostkowego. Filtr $A(z)$ można przedstawić w postaci wielomianowej za pomocą jego współczynników predykcyjnych:

$$A(z) = \sum_{i=0}^{P} a_i z^{-i} \quad a_0 = 1;\ a_P \neq 0 \tag{3.5}$$

gdzie P jest rzędem filtra. Po zastosowaniu przekształceń powyższego równania uzyskuje się zależność między współczynnikami cepstralnymi a współczynnikami predykcji. Współczynniki predykcji a_k i współczynniki cepstralne są powiązane rekurencyjnie:

$$c_1 = -a_1 \tag{3.6}$$

$$c_n = -a_n - \frac{1}{n}\sum_{i=1}^{n-1} i c_i a_{n-i} \qquad 1 < n \leq P \tag{3.7}$$

gdzie P jest rzędem filtra, a c_n jest n-tym współczynnikiem cepstralnym. Innym sposobem jest obliczanie współczynników cepstralnych za pomocą pierwiastków filtra predykcyjnego.

Dla każdej ramki sygnału definiuje się funkcję przejścia:

$$S(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{i=1}^{P} a_i z^{-i}} \tag{3.8}$$

będącą filtrem typu all-pole rzędu P, którego pierwiastki z_i filtra $A(z)$ są biegunami filtra typu all-pole. Każdy pierwiastek z_i posiada powiązane pasmo B_i oraz częstotliwość środkową ω_i dane równaniami:

$$z_i = e^{-B_i + j\omega_i} \tag{3.9}$$

$$\omega_i = \frac{1}{2\pi} arctan \frac{\Im(z_i)}{\Re(z_i)} \tag{3.10}$$

$$B_i = -\frac{1}{\pi} \ln(|z_i|) \tag{3.11}$$

Współczynniki cepstralne można przedstawić jako:

$$c_n = \frac{1}{n}\sum_{i=1}^{P} {z_i}^n = \frac{1}{n}\sum_{i=1}^{P} e^{-n(B_i + j\omega_i)} = \frac{1}{n}\sum_{i=1}^{P} e^{-nB_i} \cos(n\omega_i) \tag{3.12}$$

W [1] zbadano wpływ poszczególnych biegunów na średnią cepstralną. Z obserwacji wynika, że na bieguny charakteryzujące się węższym pasmem bliżej związane są z sygnałem mowy niż z kanałem transmisyjnym. Zaproponowana technika

filtracji biegunowej polega na poszerzeniu pasma biegunów, przy jednoczesnym zachowaniu ich częstotliwości środkowej.

Zastosowany algorytm wyznaczania charakterystyki kanału transmisyjnego:

- dla każdej ramki sygnału mowy:

  - obliczany jest wielomian współczynników predykcji liniowej A

  - obliczane są pierwiastki z_i wielomianu A

  - dla ustalonego progu α:

    - jeśli $|z_i| > \alpha$ to Z_i przypisywana jest wartość α

    - w przeciwnym wypadku $Z_i = z_i$

  - na podstawie pierwiastków Z_i obliczane są współczynniki cepstralne PFCC (*pole-filtered cepstral coefficients*) [2]

  - na podstawie pierwiastków z_i obliczane są współczynniki cepstralne LPCC (*linear prediction derived cepstral coefficients*) [1]

  - na podstawie PFCC oraz LPCC obliczane są średnie cepstralne CMN

  - współczynniki cepstralne uzyskane z CMN przeliczane są na współczynniki filtrów CCF

Filtr kompensacji kanału CCF (*Channel Compensation Filter*) obliczany jest wg. wzoru:

$$a_n = -c_n - \frac{1}{n}\sum_{i=1}^{n-1} c_{n-i}a_i \qquad 1 < n \leq P \tag{3.13}$$

jest to filtr odwrotny przeprowadzający dekonwolucję kanału transmisyjnego od oryginalnego sygnału, a zarazem zestaw parametrów opisujących kanał transmisyjny.

Znormalizowany sygnał opisany jest wzorem:

$$S_N(z) = \frac{CCF(z)}{A(z)} \tag{3.14}$$

3.2. Rozpoznanie z wykorzystaniem klasyfikatora formantów.

Obiecującą metodą rozróżnienia nagrania od rzeczywistego mówcy jest klasyfikacja na podstawie współczynników cepstralnych. Posiadając odpowiednio dużą bazę nagrań mowy oraz nagrań mowy odtworzonej (np. z głośnika) można wytrenować system tak, by potrafił rozróżniać mowę odtwarzaną od rzeczywistej. Nagrania wykorzystane w poniższym eksperymencie zostały wykonane po stronie serwera, po przejściu przez cyfrową linię telefoniczną, a następnie odtworzone z głośników i ponownie nagrane na serwerze. Każde nagranie charakteryzuje się czasem trwania 4s, częstotliwością próbkowania 8kHz oraz jakością 16bit na próbkę. Eksperyment przeprowadzono w celu sprawdzenia przydatności rozwiązania w detekcji nagrań. Do wykonania testu utworzono bazę danych złożoną z parametrów PFCC wyznaczonych dla każdej ramki, każdego pliku. Bazę podzielono na klasy nagrań rzeczywistych (1) oraz odtworzonych (2).

W celu zbadania skuteczności rozwiązania zastosowano test leave-one-out. Każdy komplet nagrań (rzeczywiste i odtworzone) został wyjęty z bazy, a następnie sklasyfikowany przy wykorzystaniu pozostałych elementów bazy. Przyjęto klasyfikator k-najbliższych sąsiadów (kNN) minimalno-odległościowy [16]. Test przebiegał wg. następującego algorytmu:

- dla pliku testowego wyznaczono średnią cepstralną ze wszystkich jego ramek (w celu przyspieszenia obliczeń)
- dla wektora parametrów wyznaczono wektory jego odległości euklidesowych od wszystkich ramek plików z klas (1) i (2)
- wektory odległości posortowano w kolejności rosnącej, a następnie uśredniono jego k pierwszych elementów
- plik przypisano do klasy, dla której średnia odległości była najmniejsza

3.3. Nakładanie się kanałów transmisyjnych.

Wykonując nagranie, które będzie odtwarzane w poczcie głosowej, użytkownik łączy się z centralą. W nagraniu zawierają się również cechy kanału transmisyjnego, na którym odbyło się połączenie przy nagraniu. W przypadku odtworzenia nagrania, przy połączeniu z pocztą głosową nakładają się zatem cechy dwóch różnych kanałów transmisyjnych. W systemach niewymagających natychmiastowej reakcji systemu (po zebraniu tylko jednej próbki mowy), jak np. w systemie weryfikacji tożsamości w telefonicznej obsłudze konta bankowego, można wykorzystać efekt nakładania się kanałów jako dodatkowy etap weryfikacji. Przykładowy proces weryfikacji:

- użytkownik kilkakrotnie proszony jest o podanie hasła (założeniem jest, że użytkownik wykonał nagrania dla bazy danych podczas poprzednich połączeń).
- uzyskane próbki mowy są przetwarzane, a następnie dokonywana jest ekstrakcja cech kanałów transmisyjnych
- następuje porównanie cech kanałów z kolejnych wypowiedzi, gdy zostaną wykryte różnice między kanałami przekraczające wcześniej ustalony próg, system odmawia użytkownikowi dostępu, w przeciwnym wypadku następuje dalsza weryfikacja
- uzyskane wektory cech porównywane są z wektorami obecnymi w bazie
- jeśli podobieństwo któregokolwiek wektora do któregokolwiek elementu bazy jest większe niż ustalony próg, system identyfikuje próbkę jako nagranie

W poniższym eksperymencie podjęto próbę zbadania skuteczności powyższego algorytmu dla opracowanych metod analizy kanału transmisyjnego.

Dla dwóch różnych wypowiedzi wykonanych na tym samym kanale transmisyjnym obliczono filtry kompensacji kanału CCF z filtracją biegunową. Obliczono odległość euklidesową uzyskanych wektorów cech. Na podstawie uzyskanego wyniku ustalono próg weryfikacji.

Do symulacji zakłóceń linii telefonicznej wykorzystano wbudowaną funkcję programu Simulink symulującą linię telefoniczną, z wykorzystaniem filtra FIR oraz generatora szumu białego. Wykonano symulację transmisji sygnału, a w efekcie splot zakłócenia linii telefonicznej z sygnałem mowy przetransmitowanym cyfrową linią

telefoniczną. Dla obu sygnałów utworzono wektory cech kanału transmisyjnego CCF z filtracją biegunową. Obliczono odległość euklidesową między sygnałami.

Aby sprawdzić skuteczność systemu w kontekście poczty głosowej wykonano dodatkowy eksperyment, symulujący sytuację, w której nagranie odtwarzane jest po upływie jednej sekundy od uzyskania połączenia. W tym celu dodano na początku nagrania sygnału zawierającego kanał transmisyjny sekundę ciszy i przeprowadzono symulację przejścia przez linię telefoniczną. Następnie oddzielnie obliczono cechy kanałów dla pierwszej sekundy nagrania i dla pozostałej jego części.

3.4. Rozpoznawanie mowy.

Odbierająca połączenie osoba korzysta z określonego i ograniczonego zakresu słów, odmiennego od zakresu słów wykorzystywanego przez domyślną pocztę głosową oraz nagranie użytkownika. Opierając się na tym założeniu można w systemie detekcji poczty głosowej zaimplementować moduł rozpoznawania mowy ze słownikiem opartym na statystycznie, najczęściej uzyskiwanych odpowiedziach po połączeniu. Skuteczny moduł rozpoznający znacznie ułatwiłby proces weryfikacji. Rozpoznając dwa lub nawet jedno słowo można uzyskać skuteczną metodę. Dodatkowo weryfikację można uprościć przyjmując pewne proste, logiczne założenia. W telefonicznych usługach marketingowych docelowym klientem jest osoba indywidualna, zatem wszelkie połączenia rozpoczynające się przedstawieniem się, podaniem nazwy firmy etc. można wyeliminować.

W celu sprawdzenia metody wykonano serię nagrywanych połączeń telefonicznych i na podstawie uzyskiwanych odpowiedzi stworzono zestaw najczęściej stosowanych zwrotów zarówno przy odebraniu połączenia przez żywą osobę jak i przez pocztę głosową. Nagrania i połączenia wykonano w programie Phoner.

Badanie skuteczności rozpoznania wykonano wprowadzając uzyskane z nagrań próbki mowy do programu Sarmata, stworzonemu na AGH systemowi rozpoznawania mowy dedykowanemu mowie polskiej [4].

4. Weryfikacja i wyniki.

4.1. Filtracja biegunowa.

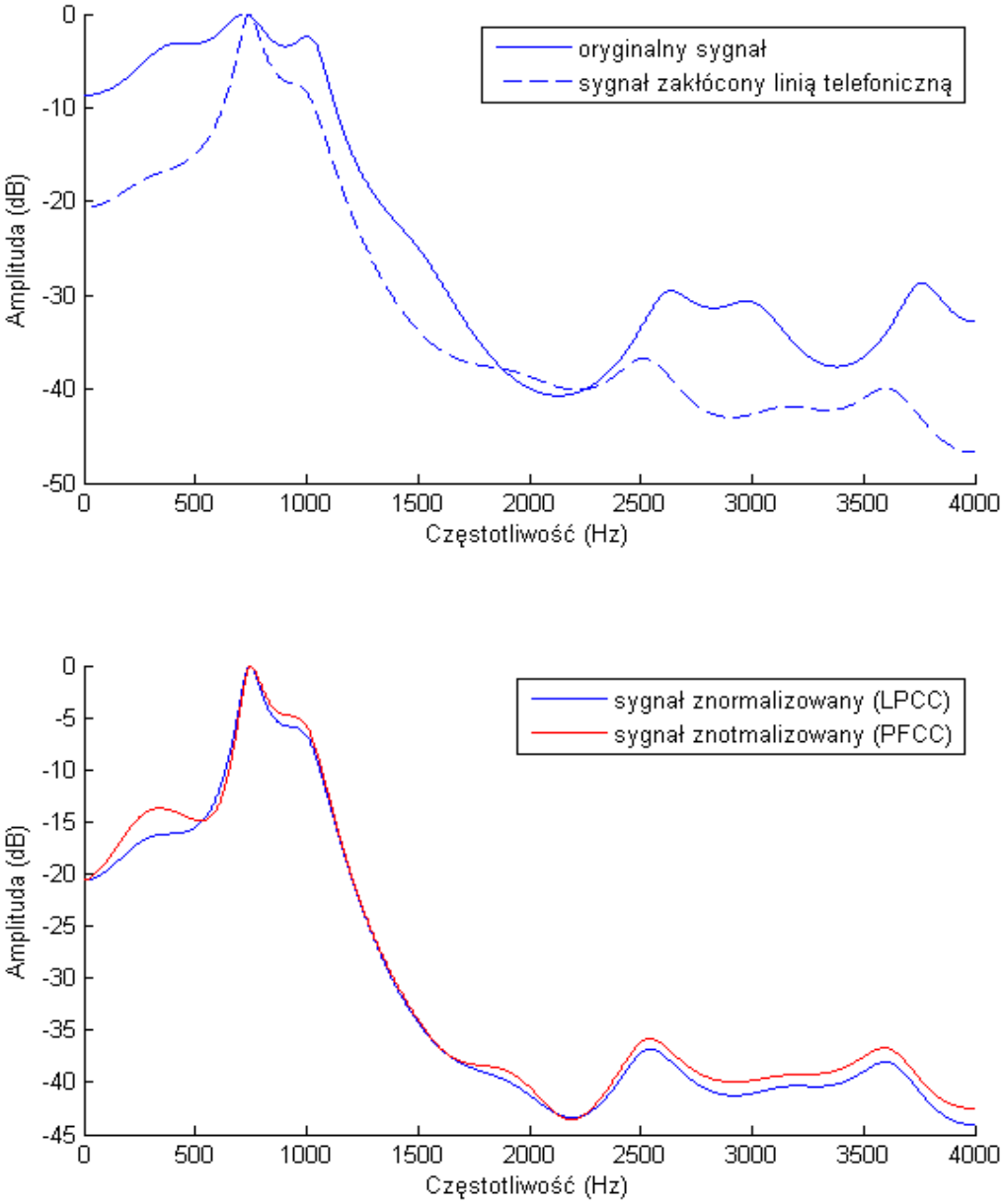
Aby zbadać skuteczność zaproponowanej metody ekstrakcji cech kanału transmisyjnego wykorzystano symulację linii telefoniczną z wykorzystaniem filtra FIR i generatora szumu białego oraz nagrania czystej mowy. Przeprowadzono splot nagrań z zakłóceniem linii telefonicznej poprzez przeprowadzenie symulacji transmisji sygnału. Zakłócone nagrania poddano normalizacji z wykorzystaniem filtrów kompensacji kanału CCF uzyskanych ze współczynników LPCC i PFCC.

Na wykresie 4.1. przedstawiono spektrum fragmentów mowy oryginalnego sygnału i zakłóconego oraz znormalizowanych różnymi filtrami CCF. Normalizacja z wykorzystaniem PFCC w porównaniu do LPCC dała niewielką różnicę w filtracji. Zaobserwować można jedynie zbliżenie sygnału znormalizowanego do pierwotnego o niewielkie wartości w dolnym i górnym zakresie pasma. Dla pozostałych ramek różnice między estymacjami zmieniały się w sposób utrudniający ocenę skuteczności rozwiązania.

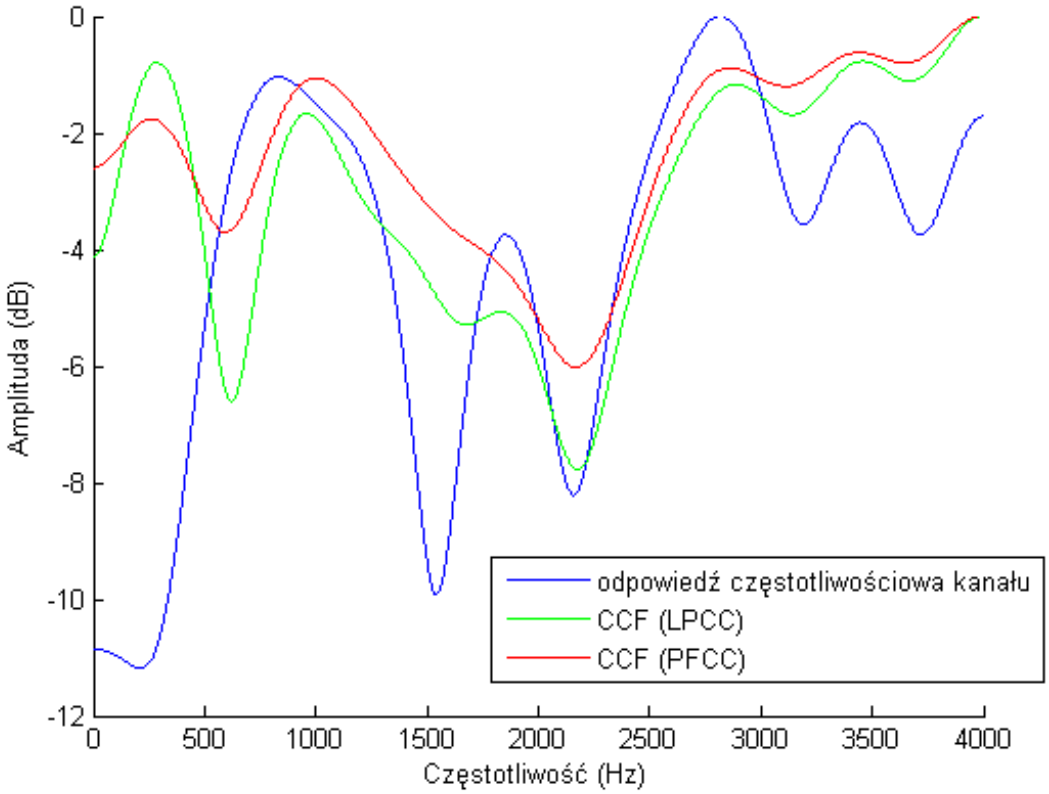
Wykres 4.2. przedstawia uzyskane estymacje kanałów. Widać wyraźne złagodzenie obwiedni spektrum kanału uzyskanego z PFCC jednak obie estymacje znacznie odbiegają od faktycznej charakterystyki kanału. Każdy z rodzajów estymacji daje bliższe oryginałowi wyniki w innych częściach pasma.

Uzyskane wyniki nie pozwalają jednoznacznie określić, czy metoda filtracji biegunowej daje poprawę estymacji kanału. Może to sugerować wybór nieodpowiedniej metody symulacji kanału transmisyjnego. Lepszą estymację za pomocą filtracji biegunowej można byłoby uzyskać stosując w implementacji ważenie poszczególnych formantów [8]. Szczegółowy opis tego rozwiązania przedstawiony został w [1].

Ponieważ celem filtracji biegunowej w omawianym rozwiązaniu nie jest poprawa jakości sygnału mowy, a jak najdokładniejsza estymacja cech kanału podjęto decyzję o zastosowaniu jej w dalszych rozwiązaniach, opierając się na poprawie wyników uzyskiwanych dzięki tej metodzie w [1].



Rys. 4.1. Zestawienie spektrum dla ramki sygnału.



Rys. 4.2. Odpowiedź częstotliwościowa zastosowanego kanału oraz estymacje kanałów metodami LPCC i PFCC.

4.2. Klasyfikator formantowy

Eksperyment przeprowadzono dla wartości $k = 5, 10$ oraz 20. Wyniki eksperymentu przedstawiono w tabeli 4.1.

Tab. 4.1

k	skuteczność rozpoznania
5	87,65%
10	90,74%
20	90,74%

Uzyskane wyniki potwierdzają skuteczność rozwiązania. Wartość rozpoznania jest zadowalająca pomimo zastosowania stosunkowo prostego klasyfikatora i małej ilości danych treningowych. Wadą wybranego rozwiązania jest duża złożoność

obliczeniowa i czas trwania obliczeń. Szybkość i skuteczność rozpoznania można znacznie poprawić stosując inne klasyfikatory (np. GMM, NTN), większą ilość danych treningowych, dłuższe nagrania treningowe (10s i więcej), optymalizację kodu oraz przetwarzanie danych wejściowych (filtracja, wycięcie ciszy etc.).

4.3. Nakładanie się kanałów transmisyjnych.

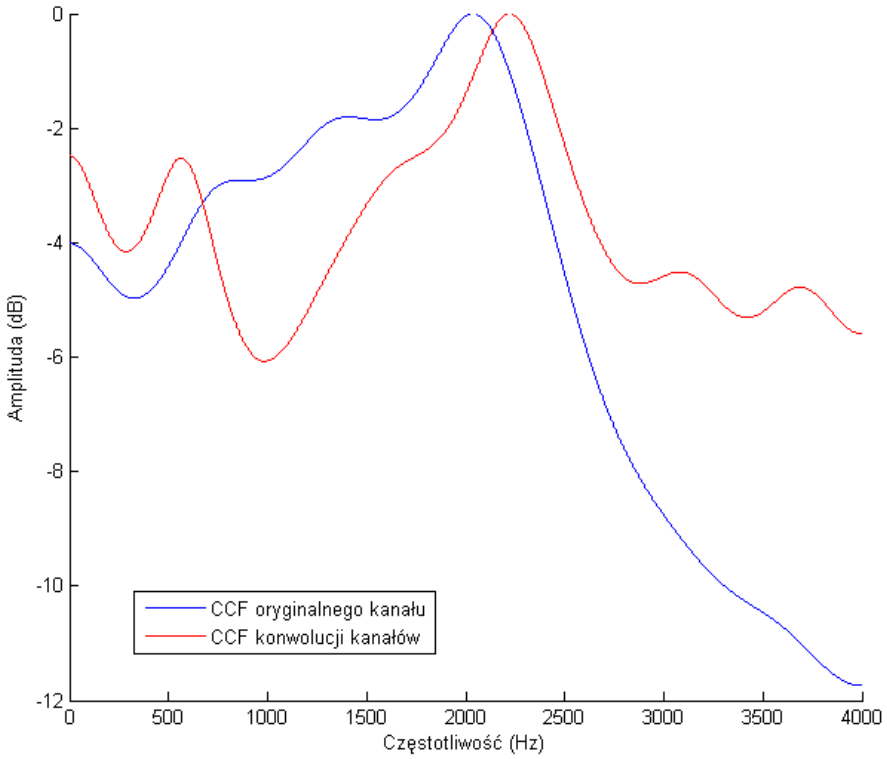
Efektem szumu białego w zastosowanej symulacji linii telefonicznej jest stosunkowo wyrównana odpowiedź impulsowa wygenerowanego kanału. Efektem jest znaczna rozbieżność z kanałem telefonii cyfrowej, zawartym w oryginalnym sygnale, a co za tym idzie, detekcja różnicy między kanałami i odmówienie dostępu użytkownikowi (wykres 4.3). Znaczna różnica między estymacjami w górnej części pasma może sugerować zły dobór parametrów symulowanego kanału.

W praktyce różnice między kanałami są znacznie mniejsze, zatem przy stosowaniu powyższej metody konieczne jest empiryczne ustalenie progu weryfikacji. Dla ustaleniu progu weryfikacji można wykorzystać różnice między dwoma różnymi wypowiedziami na jednym kanale transmisyjnym (wykres 4.4). Przy ustalaniu wartości progu kluczowe mogą okazać się również takie czynniki jak wygoda użytkownika (tolerancja na błędną weryfikację), wymagany poziom zabezpieczeń lub warunki sprzętowe.

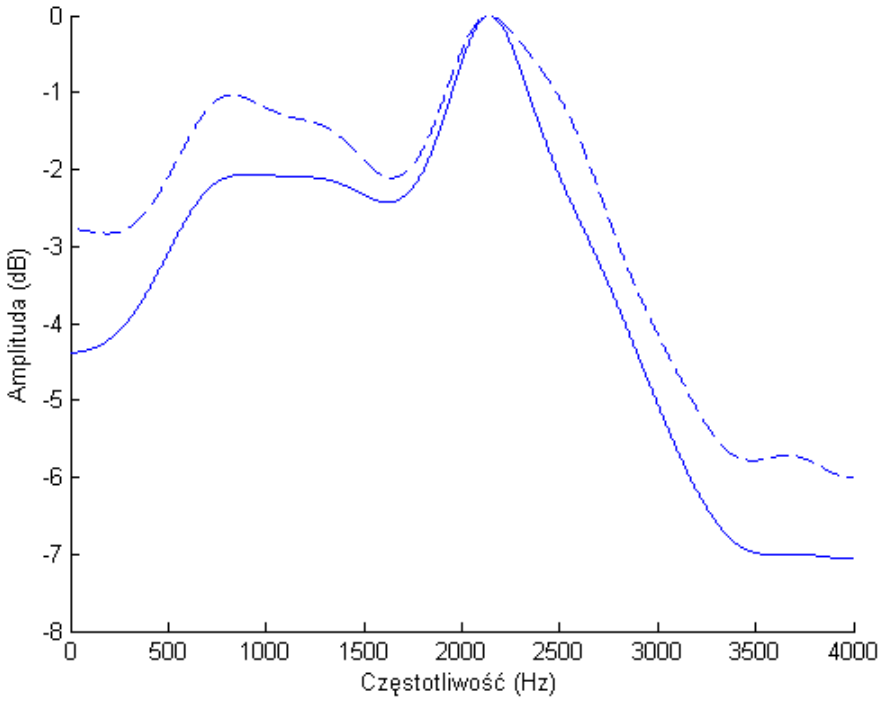
W pierwszej części eksperymentu estymacja cech kanałów przebiegała dla całego czasu trwania próbki dźwięku. W celu analizy przypadku w kontekście detekcji poczty głosowej w drugiej części eksperymentu dodano na początku nagrania sekundę ciszy, nałożono zakłócenie i dokonano ekstrakcji cech kanału dla pierwszej sekundy i pozostałej części nagrania.

Charakterystyka kanału uzyskiwana dla pierwszej sekundy zaszumionego sygnału ma różnice wartości rzędu 20dB. Trudno ocenić czy jest to efekt problemów ze skutecznością obliczeń przy krótkim czasie trwania sygnału czy też estymacja jest poprawna, a charakterystyka krótszego sygnału wynika z filtracji sygnału o wartościach zerowych. Dokładna weryfikacja metody wymaga przeprowadzenia testów z wykorzystaniem nagrań rzeczywistej sytuacji omawianej powyżej. Jedyne nagranie spośród wykonanych na potrzeby rozpoznawania mowy, zawierające autentyczne

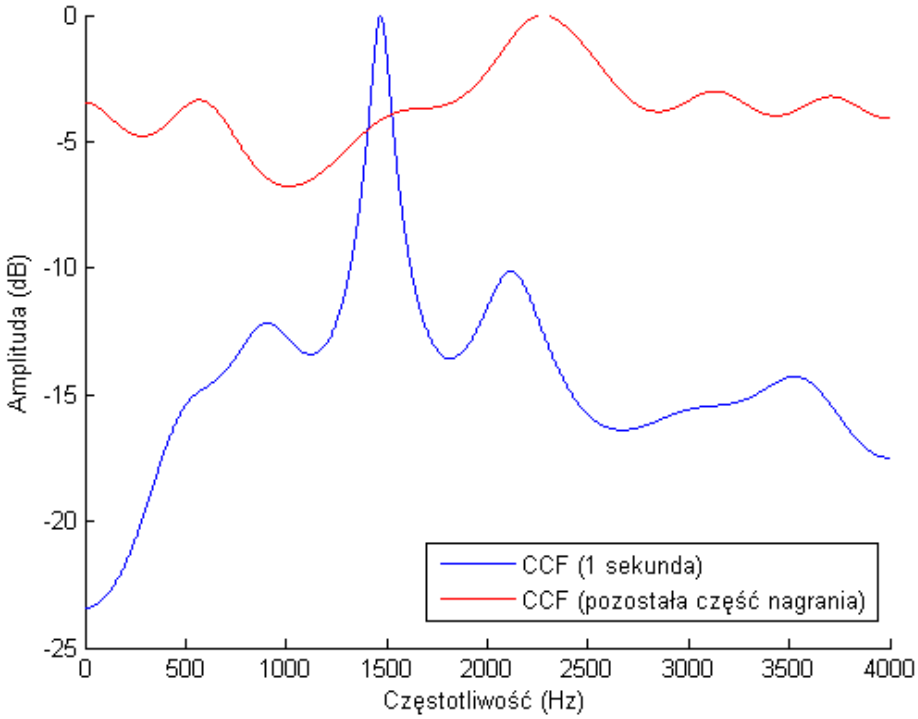
nagranie głosu użytkownika odtwarzane przez pocztę głosową, ze względu na opóźnienie między uzyskaniem połączenia a rozpoczęciem nagrywania przez program Phoner, nie posiada wymaganego do testowania metody fragmentu sygnału sprzed rozpoczęcia odtwarzania nagrania.



Rys. 4.3. Odpowiedzi impulsowe filtrów kanału oryginalnego oraz konwolucji kanałów dla jednej wypowiedzi.



Rys. 4.4. Odpowiedzi impulsowe filtrów kanałów dla dwóch różnych wypowiedzi na tym samym kanale transmisyjnym.



Rys. 4.5. Odpowiedzi impulsowe filtrów kanału dla pierwszej sekundy i pozostałej części wydłużonego i zakłóconego nagrania.

4.4. Rozpoznawanie mowy.

Część wykonanych połączeń została odebrana przez pocztę głosową, co pozwoliło wzbogacić statystykę. Uzyskane wyniki przedstawiono w tabelach 4.2. oraz 4.3. Opierając się na wynikach można stwierdzić, że w procesie weryfikacji wystarczy tak naprawdę rozpoznawanie jednego słowa. Najczęściej uzyskiwanymi odpowiedziami okazały się być cztery słowa (halo, słucham, proszę, tak) oraz ich różne kombinacje. Jeden przypadek, w którym osoba odbierająca przedstawiła się na początku można wyeliminować dla uproszczenia procesu weryfikacji.

Do weryfikacji skuteczności metody wykorzystano program Sarmata ze słownikiem stworzonym na podstawie najczęściej występujących słów. Podczas testów największym problemem okazała się jakość i czas trwania nagrań. Program miał trudność w rozpoznaniu większości słów, rozpoznawał jedynie bardzo wyraźnie wypowiedziane przez niskie głosy męskie „halo". Poprawa skuteczności rozpoznania nastąpiła po dodaniu na początku nagrań dodatkowych 300-500ms ciszy. Analiza tabel 4.2 oraz 4.3 pokazuje, że technika jest możliwa do zastosowania, gdyż żadne z wyrażeń nie występuje w obu grupach. Można założyć zatem, że skuteczność metody zależna jest w głównej mierze od skuteczności zastosowanego systemu rozpoznawania mowy. Najlepszym rozwiązaniem problemu w rzeczywistym systemie mogłoby zatem być zastosowanie innego programu lub dostosowanie Sarmaty do przetwarzania krótkich i zniekształconych wypowiedzi. Koniecznością jest również przetwarzanie wstępne sygnału, takie jak preemfaza, redukcja szumu etc.

Tab. 4.2. Wystąpienia wyrażeń używanych przy odbieraniu połączeń.

Fraza	**ilość wystąpień**
halo	58
tak słucham	16
słucham	15
tak	4
halo słucham	2
proszę	2
proszę słucham	1
<imię> słucham	1

Tab. 4.3. Wystąpienia wyrażeń w nagraniach poczty głosowej.

Treść nagrania	**ilość wystąpień**
przepraszamy …	5
tutaj <numer>	5
orange poczta …	4
t-mobile poczta …	3
osoba, do której …	3
nagraj wiadomość	2
zadzwoniłeś na nr …	1
dodzwoniłeś się …	1

5. PODSUMOWANIE

Wykorzystując zaprezentowane metody oraz tworząc dla każdej z nich odpowiednie warunki techniczne, można stworzyć kompletny system detekcji poczty głosowej, wymagający jedynie włączenia w system główny, wykonujący połączenia. Proponowany system może zawierać wszystkie przetestowane rozwiązania z odpowiednimi poprawkami.

Filtracja biegunowa przed implementacją w systemie wymaga potwierdzenia skuteczności. Sugerowane jest przetestowanie metody większą ilością nagrań oraz próba opracowania skuteczniejszych algorytmów. Dodatkowo można połączyć filtrację biegunową z ważeniem współczynników cepstralnych [8] w celu uzyskania jeszcze dokładniejszej estymacji kanału.

Zastosowany klasyfikator w testach klasyfikator jest dokładny lecz złożony obliczeniowo, co przekłada się na dłuższy czas weryfikacji. Sugerowane jest zastosowanie dowolnego z wymienionych w poprzednim rozdziale klasyfikatorów, w celu zwiększenia szybkości obliczeń. Bardzo ważną kwestią jest zgromadzenie odpowiednio dużej ilości nagrań. Baza danych powinna być zróżnicowana – głosy męskie i damskie osób w różnym wieku, a także o różnej wysokości i brzmieniu.

Podobnie jak filtracja biegunowa, metoda wykorzystująca nakładanie się kanałów wymaga dodatkowych testów przed implementacją. Zalecane jest przeprowadzenie testów na odpowiednio dużej ilości nagrań. Konieczne jest również ustalenie właściwego progu weryfikacji poprzez uwzględnienie różnorodnych czynników (np. opisanych w 4.3.). Technika polegająca na porównywaniu cech kanału przed i w trakcie trwania wypowiedzi może być celem dalszych badań, gdyż brak możliwości technicznych nie pozwolił na zbadanie jej, wystarczające do potwierdzenia jej skuteczności i z pewnością nie został sprawdzony pełen potencjał tej metody.

Weryfikacja poprzez rozpoznanie mowy, po zastosowaniu odpowiednich zabiegów, takich jak dokładne przetwarzanie danych, dostosowanie modułu rozpoznającego do warunków technicznych systemu (wzięcie w module poprawek, na np. zakłócenia w wykorzystywanej sieci telekomunikacyjnej etc.) i innych zależnych od indywidualnych preferencji użytkownika, może okazać się najsilniejszą metodą w systemie. Skuteczność programów rozpoznających mowę wzrasta z roku na rok, więc z pewnością ta część systemu jest warta wykorzystania.

Przy połączeniu wielu z proponowanych rozwiązań istotną kwestią jest dobranie odpowiedniego modułu decyzyjnego – szeregowego lub równoległego. W obu przypadkach wybór zależny jest od średniej skuteczności uzyskiwanej przez każdą z przyjętych technik. Sugerowanym rozwiązaniem jest moduł równoległy, w którym wyniki z każdej metody uzyskiwane są oddzielnie, a na podstawie uzyskiwanej w testach skuteczności, przetwarzane są przez moduł decyzyjny z odpowiednimi wagami.

Najistotniejszą kwestią przy podejmowaniu decyzji o elementach i parametrach systemu będą posiadane warunki techniczne i żądana wydajność.

6. BIBLIOGRAFIA

[1] Naik D., Mammone R.: *Communications Channels Normalizattion Techniques*, New York, Rome Laboratory 1995

[2] Naik D., Pole-filtered cepstral mean subtraction. *Acoustics, Speech, and Signal Processing*, Detroit, USA, 9-12.05.1995, 157 - 160

[3] Sharma M., Mammone R.: *System and Method for Detecting a Recorded Voice*, United States Patent 2002

[4] Ziółko M., Gałka J., Ziółko B., Jadczyk T., Skurzok D., Mąsior M., *Automatic Speech Recognition System Dedicated for Polish*, Florencja, Show and tell session, Interspeech 2011

[5] Tadusiewicz R., *Sygnał Mowy*, Warszawa, Wydawnictwa Komunikacji i Łączności, 1998

[6] Mahadeva Prasanna S. R., Cheedella S. Gupta, Yegnanarayana B., *Extraction of speaker-specific excitation information from linear prediction residual of speech, Speech Communication* 48(10), 2006, 1243-1261

[7] Parametryzacja sygnału mowy. Perceptualne skale częstotliwości, http://sound.eti.pg.gda.pl/student/amowy/AM_04_parametryzacja.pdf (odwiedzona 22.01.2014)

[8] Paliwal. K. K., *On the performance of the quefrency-weighted cepstral coefficients in vowel recognition, Speech Communication* 1, 1982, 151-154

[9] Young S., Kershaw D., Odell J., Ollason D., Valtchev V., Woodland P., *The HTK Book*, UK: Cambridge University Engineering Department, 2005

[10] Szabatin J., *Podstawy teorii sygnałów*. Warszawa, WK, 2000

[11] Kecman V., *Learning and Soft Computing — Support Vector Machines, Neural Networks, Fuzzy Logic Systems*, The MIT Press, Cambridge, MA, 2001

[12] Gaussian Mixture Models, http://scikit-learn.org/stable/modules/mixture.html, (odwiedzona 22.01.2014)

[13] Ciota Z., *Metody przetwarzania sygnałów akustycznych w komputerowej analizie mowy*, Warszawa, Exit, 2010

[14] Lyons R. G., *Wprowadzenie do cyfrowego przetwarzania sygnałów*, Warszawa, WK, 2000.

[15] Deller Jr. J. R., Hansen J. H. L., Proakis J. G., *Discrete-Time Processing of Speech Signals*, New York, IEEE, 2000

[16] Schematy testowania klasyfikatorów. Algorytm k najbliższych sąsiadów. http://edu.pjwstk.edu.pl/wyklady/adn/scb/rW9.htm, (odwiedzono 15.01.2014)