POLITECHNIKA WROCŁAWSKA
WYDZIAŁ ELEKTRONIKI

KIERUNEK: Elektronika (EKA)
SPECJALNOŚĆ: Aparatura elektroniczna (EAE)

PRACA DYPLOMOWA MAGISTERSKA

System weryfikacji mówcy w czasie rzeczywistym

Real-time speaker verification system

AUTOR:
inż. Adam Matusiak

PROWADZĄCY PRACĘ:
dr hab. inż. Józef Borkowski Prof. PWr

OCENA PRACY:

WROCŁAW 2017

Pracę ów dedukuję z serca całego psu.

Rozdział 1

Wprowadzenie

%TODO biometryk nie da się zgubić choć część populacji jest niema. Istniejąca infrastruktura pomaga. Łatwość uzyskania sygnału mowy, nawet bez kooperacji.

1.1 Weryfikacja mówcy

Weryfikacja mówcy (*speaker verification*).

Proces weryfikacji mówcy jest związany z szerszym zagadnieniem - rozpoznawania mówcy (*speaker recognition*), które charakteryzuje ogół metod wykorzystujących dane biometryczne zawarte w sygnale mowy w celu określenia tożsamości.

Sygnał mowy może być rozpatrywany jako cecha biometryczna. Sygnał mowy charakteryzowany jest przez budowę aparatu głosowego człowieka, która jest mniej lub bardziej unikatowa dla każdego człowieka, umożliwiając rozróżnienie badanej jednostki na tle populacji.

Ogólną strukturę problemu rozpoznawania mówcy można rozłożyć na na trzy elementy.[**fosr**] Po pierwsze, konieczne jest aby tworzony system dysponował modelem charakterystyk aparatu głosowego człowieka. Model taki dla przykładu może przybrać formę modelu fizyko-matematycznego aparatu głosowego człowieka. Otrzymany model musi umożliwiać parametryzację - skojarzenie z konkretną osobą. Model taki tworzony jest poprzez analizę sygnału mowy. Dopiero na tej podstawie możliwe jest porównywanie modelu utworzonego przy użyciu testowanego sygnału z modelem odniesienia. Forma i cel tego porównania definiują podklasę problemu rozpoznawania mówcy.

Weryfikacja mówcy charakteryzuje się wykonaniem dwóch kluczowych porównań - pierwszego pomiędzy modelem utworzonym z poddanego weryfikacji sygnału mowy a pamiętanym modelem osoby której dotyczy weryfikacja. W odróżnieniu of problemu identyfikacji, podczas weryfikacji mówcy potrzebna jest więc znajomość tożsamości osoby poddanej weryfikacji. Drugie z kolei porównanie dokonywane jest pomiędzy modelem poddanym weryfikacji, a uogólnionym modelem całej populacji (*background model*) lub pewnej jej podgrupy (*cohort model*. Na podstawie relacji tych dwóch odległość podejmowana jest decyzja o autoryzacji.

W przypadku kiedy nie jest możliwa lub pożądana znajomość przez system tożsamości osoby weryfikowanej przed dokonaniem autoryzacji, możliwe jest zastosowanie bardziej złożonego problem identyfikacji mówcy na otwartym zbiorze.(*open-set speaker identification*). Proces ten można uważać jako złożenie problemu weryfikacji mówcy oraz identyfikacji mówcy na zbiorze zamkniętym (*close-set speaker identification*). Polega on na przeprowadzeniu weryfikacji mówcy na modelu uzyskanym z procesu identyfikacji mówcy

na zbiorze zamkniętym, która dokonuje porównania z całą dostępną bazą modeli mówców i zwraca ten najbliższy modelowi testowanemu. Problem taki jest więc obliczeniowo co najmniej tak złożony jak weryfikacja mówcy (dla bazy w której znajduje się tylko jeden mówca).

1.2 relacja pomiędzy rozpoznawanie mowy, a rozpoznawaniem mówcy

Kluczowe jest odróżnienie procesu rozpoznawania mówcy od systemów rozpoznawania mowy (*speech recognition*). Pomiędzy tymi dwoma rozpatrywanymi dziedzinami z zakresu analizy sygnału mowy występuje dychotomiczny podział. Wynika to z tego, że sygnał mowy jest sygnałem bogatym informacyjnie oraz że jedynie mała część tej informacji posiada znaczenie semantyczne, zaś reszta niesie wiedzę o budowie konkretnego, ludzkiego narządu mowy. W problemie rozpoznawania mowy nie jest istotna tożsamość osoby wypowiadającej się, a jedynie sens jej wypowiedzi. Zatem reszta sygnału nie zawierająca odczytywanej wiadomości jest redundantna z punktu widzenia tego zagadnienia - cała informacja biometryczna jest niewykorzystywana, co za tym idzie często filtrowana przez zaimplementowany system. Z drugiej strony, w systemach rozpoznawania mówcy, w samym sednie jego zainteresowania, abstrahuje się od treści mowy. Stanowi ona jedynie środek dla dostarczenia informacji o fizjologii aparatu mowy. Dlatego prawdopodobnie system rozpoznawania mówcy usunie treść mowy, a utworzy jedynie model aparatu głosowego. Usprawiedliwia to twierdzenie o rozłączności tych dziedzin ze względu na zainteresowanie informacją zawartą w sygnale mowy.

Okazuje się, że wspomniana wyżej zależność powoduje to, że techniki przetwarzania sygnału stosowane przy analizie obu dziedzin są w zasadzie bardzo podobne.

W przypadku rozpoznawania mówcy zależnego od wypowiadanego tekstu czy rozpoznawania mówcy z generowanym tekstem informacja semantyczna wykorzystywana jest jedynie do określenia zakresu badanych głosek czy zapobieganiu problemowi żywotności. Informacja ta zatem nie wpływa na postać stosowanych technik rozpoznawania mówcy, a jedynie na optymalny ich dobór - ujawnia kontekst użycia. Innym przykładem tego typu jest zastosowania technik rozpoznawania treści języka naturalnego m. in. w odmianach omawianych systemów opartych na nagromadzonej wiedzy (*knowledge-based systems*), których zadaniem jest jedynie wzmocnienie procesu weryfikacji oraz zapobieganie wystąpienia problemu żywotności (*liveness issue*).

1.3 Klasyfikacja weryfikacji mówcy ...?

text-dependent speaker recognition [**fosr**]

Problem żywotności polega na możliwości oszukania działającego systemu weryfikacji mówcy poprzez dostarczenie na wejście takiego systemu spreparowany sygnał mowy - na przykład wysokiej jakości nagranie weryfikowanego mówcy, edytowane w odpowiedni sposób. Wraz z rozwojem technik audio zmylenie systemu niezabezpieczonego ze względu na ten typ ataku staje się coraz łatwiejsze.

text-independent speaker recognition [**fosr**]

text-prompted speaker recognition [**fosr**]

Implementacja systemu weryfikacji mówcy jest nazywana automatycznym systemem rozpoznawania mówcy (*automatic speaker verification system*).

%TODO zastosowania

%TODO system weryfikacji mówcy może być skojarzony z innymi systemami rozpoznawania biometryk

%TODO opis zastosowania weryfikacji mówcy w systemie wbudowanym

1.4 SYGNAŁ MOWY

Człowiek dysponuje doskonałymi narzędziami do przeprowadzenia procesów rozpoznawania mowy oraz rozpoznawania mówcy. Analiza oraz zrozumienie mechanizmów powstawania mowy u człowieka dostarcza podstaw do sformułowania metod syntezy języka naturalnego. Podobna analiza systemu percepcji mowy na który składa się aparat słuchowy oraz układ nerwowy związany z dekodowaniem sygnału mowy daje podstawy do identyfikacji cech którymi posługuje się ludzki organizm do efektywnego rozpoznawania mówcy.

Układ produkcji mowy oraz jej percepcji są ze sobą nierozerwalnie związane. Sposób ekstrakcji informacji przez narząd słuchu odzwierciedla fizjologię produkcji mowy - zatem może wskazać najważniejsze cechu sygnału mowy dla naturalnego procesu rozpoznania mówcy. Dlatego wydaje się właściwe prześledzenie związków pomiędzy tymi dwoma elementami.

Ludzki układ percepcji dokonuje rozróżnienia sygnałów audio poprzez rozróżnienie trzech własności: wysokości dźwięku, głośności oraz barwy dźwięku - tembru.

1.4.1 Produkcja sygnału mowy.

Aparat mowy człowieka.

Tembr głosu mówcy ustalony jest przez budowę jego dróg głosowych.

1.4.2 Percepcja sygnału mowy przez człowieka

Aparat słuchowy.

Narząd słuchu człowieka można rozpatrywać jako transduktor, co znaczy, że mapuje zmiany ciśnienia akustycznego w powietrzu na sygnał elektryczny w układzie nerwowym. Na samym początku toru przetwarzania sygnału audio znajduje się małżowina uszna, której zadaniem jest skupienie dźwięku. Sygnał akustyczny wpadający do kanału słuchowego jest filtrowany ze względu na jego fizyczne rozmiary, usuwane są niskie częstotliwości. Zmiany ciśnienia akustycznego zamieniane są na fale mechaniczne w ciele stałym na błonie bębenkowej, a następnie wzmacniane przez układ kosteczek słuchowych - młoteczka, kowadełka i strzemiączka. Strzemiączko łączy się z uchem wewnętrznym poprzez błonę okienka owalnego(łac. *fenestra vestibuli*), którego zadaniem jest wzbudzenie drgań

Układ nerwowy

1.4.3 Lingwistyka a rozpoznawania mówcy.

Występuje silny związek pomiędzy językiem a procesem rozpoznawania mówcy. Obszarami lingwistyki, które szczególnie dotyczą badanej kwestii są: fonetyka, fonologia oraz prozodia.

Fonetyka

Fonetyka zajmuje się badaniem dźwięków produkowanych przez aparat mowy człowieka. Elementarnym dźwiękiem rozpatrywanym przez fonetykę jest głoska. Z punktu widzenia całej lingwistyki jest to najmniejszy segment mowy. Budowa i funkcjonalność narządu mowy determinują zakres produkowanych głosek. Fonetyka bada podstawowe dźwięki mowy bez rozróżnianie ze względu na konkretny język czy znaczenie głoski. Fonem jest najmniejszą jednostką mowy na podstawie której możliwa jest interpretacja jej znaczenia. To znaczy, że jest semantycznie istotna. Fonem rozpatruje się ze względu na znaczenie w konkretnym języku. Głoska może być realizacją fonemu. Dwie różne głoski mogą stanowić realizację tego samego fonemu - to znaczy nieść tę samą informację semantycznie. Dla przykładu Zbiór takich głosek nazywany jest alofonem.

Z punktu widzenia mechanizmów powstawania dźwięków w ludzkim narządzie mowy, można wyróżnić trzy z których zbudowany jest każdy emitowany dźwięk mowy. Składa się na nie: - dźwięk rezonujący powstały w wibrującym źródle (np. drgające fałdy głosowe) i rezonujący w przestrzeni rezonansowej na którą składają się drogi oddechowe znajdującą się powyżej krtani, - dźwięk powstały przez nielaminarny przepływ powietrza, - dźwięk impulsowy powstały przez energiczne wypuszczenie powietrza z układu oddechowego.

Fonologia

Fonologia bada

Prozodia

Prozodia zajmuje się brzmieniowymi właściwościami mowy na które składają się trzy elementy - intonacja, akcent oraz iloczas.

1.5 Ekstrakcja cech z sygnału mowy.

Celem ekstrakcji cech z sygnału mowy (*feature extraction*) jest uzyskanie zbioru cech charakteryzujących sygnał ludzkiej mowy za pomocą technik cyfrowego przetwarzania sygnału. Jednocześnie jest to zamiana sygnału z którym zawarta jest redundantna informacja na sygnał o niskiej zawartości informacji znaczących dla problemu rozpoznawania mówcy/mowy.

Tak jak zostało wspomniane wcześniej, problem rozpoznawania mówcy w standardowym podejściu rozpatrywany jest jako problem estymacji parametrów ustalonego modelu. W zależności od rozpatrywanej odmiany rozpoznawania mówcy dokonywane są odpowiednie założenia konkretyzujące postać problemu.

W ogólności system produkcji mowy człowieka można przedstawić jako układ regulacji automatyki z kontrolerem reprezentującym układ nerwowy wraz z kontrolowaną przezeń motoryką aparatu mowy. Sygnał sterujący kontrolera powoduje pobudzenie układu charakterystyk sygnału mowy charakteryzowany przez budowę dróg głosowych. Dopiero wyjściem tego ostatniego jest akustyczny sygnał mowy. Podczas gdy problem rozpoznawania mowy próbuje wyeliminować wpływ na formułowany model postaci układu charakterystyk głosowych, tak rozpoznawanie niezależne od tekstu próbuje ustalić model charakterystyk głosowych bez względu na postać układu kontrolera. Ponieważ oba rozpatrywane modele są mocno nieliniowe najtrudniejszym zadaniem w rozpoznawaniu mówcy jest ich rozdzielenie. Z tego właśnie powodu problem ze znanym wypowiadanym tekstem jest o

wiele prostszym zagadnieniem od problemu rozpoznawania mówcy niezależnego od wypowiadanego tekstu. W tym pierwszym przypadku dokonujemy redukcji do tylko znanych pobudzeń oraz konkretyzować model kontrolera Gc.

Przedstawione w tym podrozdziale techniki ekstrakcji cech okazują się równie przydatne i powszechnie stosowane zarówna dla technik rozpoznawania mówcy jak i rozpoznawania mowy. Najpopularniejszą obecnie stosowaną metodą ekstrakcji cech z sygnału mowy jest współczynnikami cepstrum w dziedzinie częstotliwości Mela - MFCC (*mel-frequency cepstral coeficcients*).

Ważnym elementem toru przetwarzania systemu są również elementy związane z zaszumieniem sygnału mowy, przetwornikiem elektroakustycznym oraz procesem dyskretyzowania elektrycznego sygnału mowy przez przetwornik analogowo-cyfrowy. Jednak ze względu na prostotę te zagadnienia zostaną omówione oddzielnie od problemu ekstrakcji cech.

1.5.1 MFCC.

Ekstrakcja cech za pomocą współczynników cepstrum w dziedzinie częstotliwości Mela składa się z dwóch głównych etapów: uzyskania współczynników mocy w dziedzinie częstotliwości Mela na podstawie widma mocy sygnału mowy (*freqency warping*) oraz obliczenia współczynników cepstrum na podstawie uzyskanych wcześniej współczynników mocy w skali Mela. Wynikiem przeprowadzonej operacji jest uzyskanie wektora cech jx dla każdej ramki j.

Podział sygnału na ramki.

Ze względu na własność quasi-stacjonarności sygnału mowy, aby wydobyć informację dotyczącą wypowiadanej głoski konieczne jest rozdzielenie sygnału na ramki (*framing the signal*). Liczbę próbek dla pojedynczej ramki wybiera się ze względu na konieczność uzyskania lokalnej stacjonarności - w taki sposób aby móc zbadać charakterystykę częstotliwości pojedynczej głoski. Średnia długość głoski to 80 ms. Jednak trzeba mieć na względzie, że samogłoski trwają długo w stosunku do przerw pomiędzy (trwających zwykle ok. 5 ms). Zatem aby móc uchwycić krótsze głoski oraz przerwy zwykle ustala się długość ramki na 20 do 30 ms. Jednocześnie długość ramki w ilości próbek jest funkcją częstotliwości próbkowania. Do przedstawionych założeń dochodzi zwykle warunek dotyczący użycia algorytmu szybkiej transformaty Fouriera (*Fast Fourier Transform, FFT*) wymagającej aby sygnał był długości potęgi dwójki. W przypadku pierwszej oraz ostatniej ramki stosowana jest techinka uzupełniania zerami w przypadku braku próbek do zapełnienia całej ramki.

Aplikacja okna na ramki.

W stosowanej, zredukowanej techince STFT stosowane są okna inne niż okno prostokątne o długości ramki lh w celu polepszenia właściwości widmowych dla dalszej analizy. Zastosowanie okna czasowego odbywa się poprzez przemnożenie wartości próbek ramki przez wartości okna w następujący sposób:

Najpopularniejszymi oknami stosowanymi w tym przypadku są okna czasowe Hanna, Hamming oraz okna Gaussowskie.

Estymacja widma.

Po zaaplikowaniu okien czasowych dla kolejnych ramek, kolejnym etapem do otrzymania współczynników MFCC jest znalezienie widma ramki za pomocą dyskretnej transformaty fouriera. Kolejne współczynniki uzyskanego widma zdefiniowane są następująco: W omawianej metodzie korzysta się jedynie z informacji o amplitudzie uzyskanego widma. Wartość wielkości amlitud otrzymywane są w następujący sposób:

Przejście do dziedziny częstotliwości Mela.

Przejście do dziedziny częstotliwości Mela (*frequency warping*) uzasadnione jest nieliniową percepcją układu słuchowego człowieka. Ponieważ procedura DFT produkuje liniowe widmo w dziedzinie widmo częstotliwościowe, dokonywana jest jego transformacja do dziedziny częstotliwości Mela. Rozpatrywana operacja polega na zastosowaniu banku filtrów trójkątnych na uzyskanych amplitudach widma kHz: W wyniku tej operacji otrzymuje się wektor współczynników częstotliwości w skali Mela o długości równej ilości użytych filtrów. Najczęściej stosowanym filtrem jest okno trójkątne zdefiniowane w dziedzinie częstotliwości. Każdy kolejny filtr zaczyna się w środku pasma wcześniejszego. Ponieważ przetwarzany jest sygnał składający się z próbek o wartościach rzeczywistych, pokryte pasmo częstotliwości zawiera próbki od 0-wej do próbki o oznaczonej numerem N/2. Ilość filtrów oraz ich rozmieszczenie w dziedzinie częstotliwości ustalane są na jeden z kilku sposobów. Pierwszym sposobem jest skorzystanie z 24 obszarów krytycznych zdefiniowanych w skali Barka. Każdy filtr ma niezerowe wartości w zakresie częstotliwości od k-1 częstotliwości krytycznej, aż do k+1 częstotliwości krytycznej. Maksimum zaś przypada na k-tą częstotliwość krytyczną. Drugim i najpopularniejszym podejściem jest zastosowanie ustalonej ilości filtrów i rozmieszczenie ich równomiernie w dziedzinie częstotliwości mela.

Stosowane są również filtry trapezowe oraz kosinusowe. Innym proponowanym podejściem jest znalezienie filtrów poprzez proces uczenia maszynowego, uczoną ze względu na

Cepstrum.

Funkcja cepstrum mocy dla ciągłego przekształcenia Fouriera zdefiniowana jest następująco [**fosr**]:

$$\tilde{h_{pc}} = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} log\left(|H(\omega)|^2\right)e^{i\omega t}\right]^2 \tag{1.1}$$

gdzie $|H(\omega)|^2$ oznacza funkcję widma gęstości mocy sygnału oryginalnego. Jest to odwrotne przekształcenie Fouriera zastosowane na logarytmie funkcji gęstości mocy sygnału.

Pojęcie cepstrum pierwotnie zostało wprowadzone do badań nad echem sygnałów akustycznych [**hdsp**]. Taki sygnał opisywany jest przez wyrażenie

$$x(t) = s(t) + \alpha s(t-\tau) \tag{1.2}$$

gdzie α jest współczynnikiem osłabienia sygnału echa przesuniętego w czasie o τ względem sygnału oryginalnego. Reprezentacja częstotliwościowa zaś przyjmuje w takim wypadku postać

$$|S(f)|^2\left[1+\alpha^2+2\alpha cos(2\pi f\tau)\right]. \tag{1.3}$$

Jak widać z tej postaci, widmo sygnału oryginalnego $S(f)$ modulowane jest przez zmienny w częstotliwości komponent $[1+\alpha^2+2\alpha cos(2\pi f\tau)]$. Znajomość tego w jaki sposób modulowana jest obwiednia widma pozwala na określenie współczynników α oraz τ definiujących

sygnał echa. Ponieważ problem taki jest doskonale opisany w dziedzinie czasu przy analizie sygnałów zmodulowanych, narzucającym się rozwiązniem jest skorzystanie z tych znanych już technik. Korzystając z własności logarytmu możliwe jest rozdzielenie dwóch komponentów - zamianę mnożenia na dodawanie, a poprzez zastosowanie odwrotnego przekształcenia Fouriera otrzymuje się sumę dwóch funckji w dziedzinie tzw. "quefrency":

$$\begin{aligned}\mathcal{F}^{-1}\{log(X(f))\} = \mathcal{F}^{-1}\{log(S(f))\} + \mathcal{F}^{-1}\{log(1+\alpha^2)\}+ \\ +\mathcal{F}^{-1}\{log(1+\frac{2a}{1+a^2}cos(\omega\tau))\}.\end{aligned} \tag{1.4}$$

Ostatni składnik powyższej sumy objawia się w postaci widocznego w cepstrum skupionego impulsu którego położenie w skali cepstrum określa przesunięcie τ zaś jego amplituda jest związana z czynnikiem osłabienia α.

Okazuje się, że problem wykrycia wysokości tonu (*ang. pitch detection*) sygnału mowy jest podobny do problemu analizy echa sygnału akustycznego. Zaproponowana została więc wersja wykorzystująca cepstrum dla analizy STDFT[**noll**]. Tak jak wspomniano wcześniej w tym rozdziale %TODO ODNOŚNIK prosty model produkcji sygnału mowy zakłada, że ten jest wynikiem splotu odpowiedzi impulsowej dróg głosowych h_1 oraz quasi-okresowym sygnałem impulsowym h_2 (*ang. quasi-periodic pulse train*) produkowanym w głośni:

$$h(t) = h_1(t) * h_2(t) \tag{1.5}$$

Z własności ciągłego przekształcenia Fouriera wynika, że powyższe równanie w dziedzinie częstotliwości przybiera postać iloczynu transformat $H_1(\omega)\cdot H_2(\omega)$. W dziedzinie cepstrum rozdziela się natomiast na sumę składników:

$$\tilde{h}(t) = log(H_1(\omega)) + log(H_2(\omega)) \tag{1.6}$$

Poprzez zastosowanie cepstrum sygnału mowy otrzymujemy więc rozdzielenie komponentów pochodzących od funkcji $h_1(t)$ i $h_2(t)$. Proces zobrazowany jest na rysunku %TODO RYSUNEK. Niskie współczynniki cepstrum opisują charakterysytkę aparatu głosowego, zaś widoczny w okolicach % OPISAĆ RYSUNEK.

Należy zwrócić uwagę, że wykorzystywany w tej metodzie jest logarytm z liczb rzeczywistych oraz współczynniki mocy widma. Spowodowane jest to, że podczas procesu rozpoznawania mówcy nie interesuje nas rekonstrukcja sygnału, anie nie korzysta się z informacji zachowanej w fazie sygnału. Z tego też powodu klasycznie w ostatnim etapie obliczania współczynników MFCC zamiast odwrotnej dyskretnej transformaty Fouriera (IDFT) stosuje się dyskretną transformację kosinusową (DCT). Kolejną korzyścią jest otrzymanie w wyniku przeprowadzenia tej transformacji współczynników będącymi liczbami rzeczywistymi. W tej pracy korzysta się z następującej definicji dyskretnej transformaty kosinusowej:

$$H_k = \sum_{n=0}^{N-1} h_n cos\left(\frac{\pi(2n+1)k}{2N}\right) \tag{1.7}$$

gdzie

$$a_k = \begin{cases} 1/N & \text{dla} \quad k = 0 \\ 2/N & \forall \quad k > 0. \end{cases} \tag{1.8}$$

Uzyskane współczynniki MFCC noszą nazwę wektora akustycznego (*ang. acoustic vector*) i są w procesie weryfikacji mówcy dalej wykorzystywane w tzw. procesie dopasowywania cech (*feature matching*).

1.5.2 Dopasowanie cech.

% inne techniki DTW, HMM.

Spis treści