PROCESORY SYGNAŁOWE - PROJEKT

Rozpoznawanie słów ze słownika
z wykorzystaniem LPC oraz DTW

Autorzy: Damian Dybek, Karol Draszawka.
Luty 2009.

1. Cel projektu:

Celem projektu było stworzenie na procesorze ADSP-21161 uproszczonego systemu rozpoznawania mowy. Rozpoznawanie jest zależne od mówcy, tj. działa prawidłowo dla osoby, która wcześniej przygotowała wzorce. System umożliwia przygotowanie wzorców, czyli utworzenie tzw. „słownika", poprzez nagranie wypowiedzi wyrazów; oraz rozpoznawanie, czyli wskazanie przez system, które ze słów z tego słownika zostało wypowiedziane. Ze względu na ograniczoną pamięci wewnętrznej procesora (pamięć zewnętrzna nie jest wykorzystywana) ilość słów w słowniku została ograniczona do trzech. Rozpoznanie odpowiedniego wyrazu jest sygnalizowane zaświeceniem się diody, odpowiedniej do pozycji tego słowa w słowniku, np. wykryto, że wypowiedziano słowo zapisane na pozycji drugiej w słowniku - zapali się dioda na pozycji drugiej.

2. Specyfika problemu i zastosowane rozwiązanie:

Przedstawione zadanie jest zadaniem klasyfikacji sygnału mowy. Klasyfikacja, ogólnie rzecz biorąc, dzieli się na dwa etapy, ekstrakcji cech dystynktywnych z sygnału, oraz właściwego zaklasyfikowania sygnału na podstawie tych cech. Najpierw jednak trzeba zarejestrować sygnał, który mamy zaklasyfikować.

Pozyskiwanie sygnału

Jako, że jest to sygnał mowy, dokonuję się tego za pośrednictwem mikrofonu, z którego analogowy sygnał jest zamieniany na postać cyfrową przez kodek dostępny na płytce laboratoryjnej. Ponieważ pasmo mowy nie przekracza 4000 Hz, dlatego wystarczyłoby, gdyby kodek próbkował sygnał z fs=8000Hz. Ponieważ jednak nie da się ustawić w tym kodeku takiej częstotliwości próbkowania, to sygnał jest próbkowany ze standardową fs=48000, po czym dokonywana jest decymacja - do dalszego przetwarzania brana jest co 6 próbka.

Etap ekstrakcji cech z sygnału

Sygnał mowy jest silnie zmienny w czasie. Dlatego konieczne podzielenie sygnału na odcinki występujące w krótkich ramkach czasowych trwających 20ms. Dla sygnału mowy próbkowanego z częstotliwością 8000Hz, okno takie ma długość 160 próbek. Ze względu na oszczędzanie pamięci, nie zastosowano nakładkowania ramek. Z sygnału trwającego 1 sekundę, mamy więc 50 ramek trwających 20ms. Każdą z takich ramek należy następnie poddać analizie w wyniku czego uzyskując parametry, które umożliwiałyby na podstawie nich rozpoznawanie mowy. W tym projekcie, parametrami tymi są współczynniki LPC (linear predictive coding). Obecnie częściej używanymi współczynnikami są MFCC (mel frequency cepstral coeficients), jednak ze względu na większą złożoność obliczeniową, a podobną skuteczność, zdecydowano się pozostać przy współczynnikach LPC.

Możliwe byłyby dwa podejścia:

- zapis przebiegu całego sygnału, następnie podzielenie go na ramki czasowe i dla każdej z tych ramek wyliczenie współczynników LPC, następnie ich zapis do macierzy w pamięci, na końcu usunięcie zapisanego przebiegu czasowego sygnału.
- wyliczanie współczynników LPC po każdej zarejestrowanej ramce czasowej (po każdych 20ms nagrania) i sukcesywne wypełnianie macierzy w pamięci tymi współczynnikami, nie rejestrowanie samego sygnału.

Zastosowana podejście drugie, gdyż umożliwia to zaoszczędzenie pamięci (przebieg sygnału w czasie nie jest zapisywany - jedynie pojedyncza ramka) oraz zaoszczędzenie czasu (nie ma ekstrakcji cech sygnału po nagraniu, a w trakcie nagrywania).

Wyliczanych jest 10 współczynników LPC. Odbywa się to w dwóch funkcjach. Najpierw liczone są współczynniki autokorelacji pomiędzy próbkami ramki. Następnie, z nich uzyskuje się współczynniki LPC stosując algorytm Levinsona-Durbina.

Etap właściwego rozpoznawania

Rozpoznawania polega na specyficznym porównywaniu wektora parametrów, opisujących sygnał (utrzymywany w pamięci w postaci macierzy liczb) otrzymanym w etapie poprzednim, ze zbiorem takich wektorów będących wektorami wzorcowymi - opisującymi wcześniej nagrane wypowiedzi wyrazów słownika.

Ponieważ dany wyraz można wypowiedzieć wolniej lub szybciej, proste porównanie sygnałów, ramka do ramki, nie przyniosłaby oczekiwanych efektów. Dlatego stosuje się specjalny algorytm wykorzystujący programowanie dynamiczne, tzw. dynamic time warping (DTW).

DTW jest algorytmem pozwalającym znajdować najlepsze dopasowanie pomiędzy dwoma rozciągniętymi w czasie sygnałami. Dopasowanie to polega na „zawijaniu" miejsc rozciągniętych w czasie w stosunku do wzorca i rozciąganiu miejsc, które są nazbyt skrócone - jest to znajdywanie korespondujących ze sobą obszarów pomiędzy dwiema seriami danych:

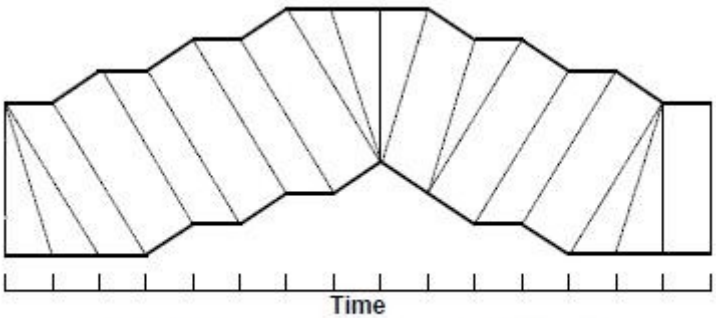
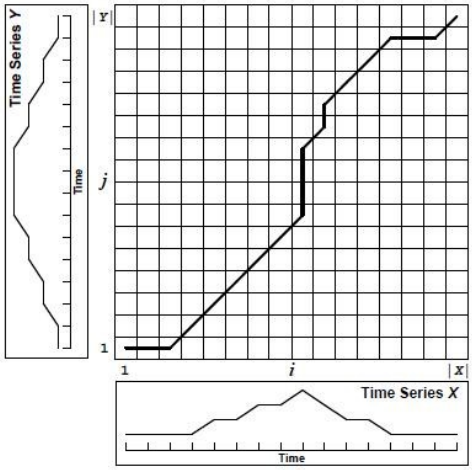


Figure 1. A warping between two time series.

Algorytm DTW znajduje optymalną ścieżkę „zawinięcia czasu”, tak aby porównywane dwa sygnały miały najmniejszą odległość względem siebie (różniły się najmniej):



W podejściu tym należy wyliczyć macierz odległości (w sensie Euklidesa) między każdą ramką sygnału badanego, a sygnału wzorca. Następnie trzeba wyliczyć macierz skumulowanej odległości między sygnałami poruszając się od elementu macierzy oznaczający odległość pierwszych ramek sygnałów, potem dodając najmniejszy z sąsiednich elementów macierzy, potem najmniejszy z elementów sąsiednich właśnie dodanego elementu itd. Powstaje ścieżka optymalnego dopasowania (jak na rysunku powyżej), otrzymujemy także minimalną odległość między sygnałem badanym, a danym wzorcem.

Należy obliczyć za pomocą DTW odległości między wszystkimi wzorcami, a badanym sygnałem. Sygnał klasyfikujemy jako ten, do wzorca którego, odległość jest minimalna. Dodatkowo, można założyć pewien próg, poniżej którego ta minimalna odległość musi zejść, by uznać, że wypowiadany sygnał, to jeden z tych, które znajdują się w bazie.

3. Użytkowanie systemu:

Stworzenie słownika, zawierającego słowa, które będą następnie rozpoznawane wymaga nagrania wzorców.

Nagranie wzorca odbywa się po naciśnięciu odpowiedniego przycisku:

Przycisk pierwszy (od lewej, w dolnym rzędzie)	Nagrywanie wzorca pierwszego	Miga dioda pierwsza (od lewej)
Przycisk drugi	Nagrywanie wzorca drugiego	Miga dioda druga
Przycisk trzeci	Nagrywanie wzorca trzeciego	Miga dioda trzecia

Każde nagranie trwa 1 sekundę.

Mając utworzony w ten sposób słownik, możemy testować rozpoznawanie wypowiadanych słów. Dokonuję się to w sposób bardzo podobny, jak tworzenie słownika. Po naciśnięciu przycisku czwartego (miga ostatnia dioda) należy powiedzieć słowo. Także i to nagranie trwa 1 sekundę. Natychmiast po wypowiedzeniu zapala się dioda, odpowiednio do pozycji w słowniku wyrazu rozpoznanego. Dioda ta pozostaje zapalona do czasu kolejnego testowania (rozpoznawania) lub nagrywania innego słowa do słownika - możliwe jest bowiem zmienianie zawartości słownika. Jeśli podczas testowania, zostało wypowiedziane słowo, którego nie ma w słowniku, nie zapala się żadna z diod.

4. Literatura:

Artykuły (dostarczone w folderze z projektem):

- „A good introduction to LPC”, dr. Sung-won Park, Texas A&M University-Kingsville.
- „FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space”, Stan Salvador and Philip Chan.
- dokumentacja zestawu uruchomieniowego 21161N EZ-KIT LITE