Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki

KATEDRA AUTOMATYKI

RAPORT

TOMASZ DRZEWIECKI

GŁOSOWE STEROWANIE ŚRODOWISKIEM GNOME 3 NA PODSTAWIE ROZPOZNAWANIA FONEMÓW Z ZASTOSOWANIEM SIECI NEURONOWYCH

Kraków 2012

Spis treści

1. Wstęp 3
2. Realizacja algorytmu z artykułu 4
2.1. Rozdzielenie fonemów 4
2.2. Wyliczenie cech 5
2.3. Rozpoznanie 5
3. Wykorzystany algorytm 6
3.1. Podział na fonemy 6
3.2. Ekstrakcja cech 6
3.3. Rozpoznanie fonemów 7
3.4. Dopasowanie do wzorca 7
4. Wyniki i wnioski 8

1. Wstęp

Celem pracy było zrealizowanie, w pełni konfigurowalnego sterowania głosowego w środowisku GNOME 3. Pod pojęciem w pełni konfigurowalnego przyjmuje się możliwość dowolnego zdefiniowania rozpoznawanych wyrazów oraz przyporządkowanych akcji. Ideałem była by możliwość wpisywania rozpoznawanych słów w sposób ?gramatyczny?. W zrealizowanym programie nie udało się tego osiągnąć i słowa wpisywane są fonetycznie.

Metoda rozpoznawania poszczególnych fonemów miała być zrealizowana w sposób opisany w ?odwołanie?. Niestety ta metoda nie dała spodziewanych rezultatów i rozpoznawanie zostało zrealizowane za pomocą autorskiego algorytmu opisanego w tej pracy.

Zawartość tej pracy to: rodział 2 przedstawia realizację algorytmu opisanego w ?odwołanie?, rodział 3 prezentuje wykorzystany algorytm, rodział 4 przedstawia wyniki oraz wnioski.

2. Realizacja algorytmu z artykułu

Algorytm jest podzielony na kilka części, które są przeprowadzane szeregowo. Są to: podział mowy na fonemy, wyliczenie cech z danego fragmentu oraz rozpoznanie. Wszystkie są opisane poniżej. Następną częścią powinno być dopasowanie znalezionych fonemów do zdefiniowanych wzorców, ale nie została zaimplementowana, ponieważ nie udało się uzyskać odpowiednich wyników w częściach wcześniejszych.

2.1. Rozdzielenie fonemów

Przy rozpoznawaniu pojedynczych fonemów pierwszą czynnością, którą należy wykonać jest podział sygnału mowy na pojedyncze fonemy. W tym celu wyliczane jest ALCR ?odowłanie?. ALCR jest to „Average Level Crossing Rate". Jest to średnia liczba przecięcia ustalonych progów przez sygnał mowy. Wykorzystany sposób wyliczenia ALCR jest następujący:

– Wyznacznie progów. Do tego celu użyto rekomendacji w ?odowłanie? dotyczącej rozmieszczenia progów. Największa gęstość progów powinna być w pobliżu zera oraz w pobliżu maksymalnej wartości. Natomiast im bliżej środka, tym bardziej gęstość progów maleje.

– Wyliczenie LCR („Level Crossing Rate"). Dla każdej próbki wyliczana jest liczba progów, które zostały przecięte przez linię łączącą daną próbkę z próbką poprzednią. W oryginalnym algorytmie ta wartość była uśredniana w pewnym otoczeniu, dla każdego progu. W implementacji ten krok został pomienięty, ponieważ znacznie zwiększał czas trwania obliczeń, a poprawa działania algorytmu znikoma, jeśli w ogóle istniałą.

– Wyliczenie ALCR. ACLR jest to średnia z wyliczonych w poprzednim kroku wartości z pewnego zakresu. W opisie algorymtu oraz w implementacji użyto otoczenia o promieniu 100 próbek.

Po wyliczeniu ALCR należy znaleźć granice fonemów. Tymi granicami są minima lokalne funkcji ALCR, które spełniają następujące warunki:

– Minimalna długość trwania fonemu to 12ms.

– Minimum lokalne musi być minimum w zakresie 20ms.

Znalezione w ten sposób minima są punktami wyznaczającymi granice fonemów.

2.2. Wyliczenie cech

Ten etap jest przeprowadzany tylko dla tych fragmentów sygnału wyznaczonych w poprzednim etapie, w których stwierdzono obecność mowy. Określanie obecności mowy odbywa się za pomocą obliczania energii fragmentu sygnału.

Wektorem cech jest wektor binarny otrzymywany z PSD („Power Spectral Density").

PSD jest obliczane jako szybka transformata Fouriera funkcji autokorelacji sygnału wejściowego. Ponieważ ustalona wielkość wektora to 513 bitów, to autokorelacji podlega jedynie 512 początkowych próbek w danym fonemie (które powinny być reprezentatywne). Po operacji autokorelacji do transformacji Fouriera używanych jest 1024 próbki. Po zastosowaniu transformacji Fouriera wybieranych jest 513 próbek, poddanych wcześniej operacji obliczania modułu.

Wyliczone w ten sposób PSD jest poddawne operacji progowania. Niestety nie przedstawiono w artykule sposobu dobierania wartość progu. Prób został przyjęty jako 20% wartości maksymalnej PSD. Wartościom powyżej progu przypisywano 1, a poniżej progu 0. W ten sposób otrzymano wektor cech.

Na tym etapie zauważono, że kształ PSD dla fonemu „s" odbiega od przedstawionego w ?odwołanie?.

2.3. Rozpoznanie

Do rozpoznawanio użyto sieci neuronowej typu „Back Propagation" o 513 wejściach oraz 34 wyjściach (po jednym dla każdego fonemu). Próby uzyskania optymalnej struktury sieci nie udały się. Probowano sieci dwu- oraz trzywarstwiwe o różnych liczbach neuronów ukrytych, ale żadna sieć osiągnęła wystarczająco niskiego błędu. Sieci o najmniejszych błędach dla danych uczących nie były w stanie rozpoznawać poprawnie fonemów, nawet z ograniczonego zbioru.

W związku z brakiem uzyskania satysfakcjonujących rezultatów próbowano zastosować inne wektory binarne np. otrzymane z transformaty Fouriera sygnału oryginalnego, bądź łączących oba te wektory, lecz żadne z rozwiązań nie dawała zadowalających efektów. Zdecydowano wybrać zupełnie inny zestaw parametrów opisujących sygnał, co jest opisane w części 3.

3. Wykorzystany algorytm

Zastosowany algorytm jest podobny do poprzedniego i składa się z następujących części: podział na fonemy, ekstrakcja cech, rozpoznanie fonemów i dopasowanie do wzorca.

3.1. Podział na fonemy

Podział na fonemy został przeprowadzony dokładnie tak samo jak w poprzednim rozwiązaniu, z użyciem ALCR. Należy tutaj jednak wspomnieć o wadach tego podziału:

- Nie zawsze dokładnie rozdziela fonemy w punkcie, który podczas analizy wzrokowej uznano by za punkt graniczny.
- Dłuższe fonemy są często dzielone na większą liczbę fragmentów.

Wspomniano o wadach, ponieważ mają one wpływ na jakość rozpoznawania oraz konieczność ich uzględnienia w następnych etapch algorytmu. Rośnie także złożoność obliczeniowa algorytmu.

3.2. Ekstrakcja cech

Do rozpoznania wybrano cechy otrzymane z postaci czasowej i częstotliwościowej. Wybrano 12 parametrów oraz zbadano ich zmienność zarówno w obrębie pojedynczych fonemów, jak i zmienność pomiędzy fonemami. Te, które określały się najlepszą zmiennością dla różnych fonemów oraz najmniejszą dla pojedynczych fonemów zostały użyte do rozpoznawania. Wszystkie zostały znormalizowane. Są to:

- Liczba przejść przez zero,
- Stosunek największej wartości w danym fragmencie do najmniejszej,
- Moment spektralny zerowego rzędu,
- Moment znormalizowany pierwszego rzędu,
- Moment scentralizowany drugiego rzędu,
- Moment scentralizowany trzeciego rzędu,
- Stosunek mocy częstotliowści wysokich do częstotliwości niskich.

Wartości czasowe są liczone bez preemfazy, a wartości częstotliwościowe są liczone z preemfazą.

3.3. Rozpoznanie fonemów

Rozpoznanie fonemów odbywa się na podstawie dwóch prawdopodieństw: prawdopodobieństwa, że dany fonem wejściowy jest określonym fonemem oraz na podstawie prawdopodobieństwa, że dany fonem powinien wystąpić w danym miejscu (obliczanym na podstawie wzorców do rozpoznania).

Wyznaczono zakresy zmienności dla każdego parametru i każdego fonemu. W ten sposób otrzymano rozkład wartości parametrów dla każdego fonemu. Z rozkładu wyznaczono wartości prawdopodobieństw dla każdego parametru i fonemu. W ten sposób otrzymując na wejściu wektor cech otrzymujemy prawdopodobieństwo wystąpienia danego fonemu.

W celu wyznaczenia prawdopodobieństwa wyliczanego z wzroców oblicza się prawdopodobieństwo wystąpienia fonemu następującego po poprzednio rozpoznanym.

Do rozpoznania używana jest sieć neuronowa. Wejściami do sieci są:

– Prawdopodobieństwa dla danego fonemu,

– Prawdopodobieństwa dla fonemu o największym prawdopodobieństwie wyliczonym za pomocą wzorców dla fonemów,

– Prawdopodobieństwa dla fonemu o największym prawdopodobieństwie wyliczonym za pomocą słów do rozpoznania,

– Wartości binarnej określającej czy poprzednio rozpoznany fonem jest badanym fonemem.

Ostatnia wartość polepsza działanie sieci dla długi fonemów, które składają się z kilku fragmentów.

Użyta sieć neuronowa jest siecią typu „Back Propagation”, z siedmiowa wejściami, trzema warstwami ukrytymi (pięć nieuronów oraz dwa neurony) oraz dwa wyjścia. Jedno wyjście określa rozpoznanie fonemu a drugie określa brak rozpoznania fonemu. Wielkość sieci została określona na podstawie badań empirycznych.

Dla jednego fragmentu rozpoznawanego może być rozpoznanych kilka fonemów.

3.4. Dopasowanie do wzorca

Dopasowanie do wzorca odbywa się na podstawie sprawdzania czy kolejne rozpoznane fonemu są dopasowane do jakiegoś wzorca. Są określone następujące warunki pozwalające na dopasowanie:

– Jeden fonem w wyrazie może zostać nie rozpoznany,

– Jednemu fonemowi we wzorcu może być dopasowanych kilka fragmentów,

– Mogą być fragmenty pominięte, z uwagi na szum oraz inne artefakty w sygnale.

Niestety w te sposób określone dopasowanie o wzorca ma dużą złożoność obliczeniową.

4. Wyniki i wnioski

Zagadnienie rozpoznawania pojedynczych fonemów jest trudne. Przedstawiony algorytm może być jedynie początkiem do dalszych prac, które mogłby usprawnić jego działania. Posiada on wiele wad oraz efekty jego działania odbiegają od ideału.

Skuteczność rozpoznawania wzorcowych słów określonych w postoci fonetycznej jest mała, ale nie jest zerowa. Podczas eksperymentów wyraz „karta" zostawał rozpoznany w 50% przypadków.

Złożoność obliczeniowa jest bardzo duża, co powoduje, że już przy kilku wyrazach czas oczekiwania przekracza kilka sekund.

W dalszych pracach polecane jest rozpoczęcie od wyszukania parametrów, które byłyby w stanie poprawniej i dokładniej klasyfikować fonemy. Poprawić należy również sam algorytm dopasowania wzorców, ale ponieważ jest on ostatnim elementem przetwarzania powinno być opracowane na końcu w przypadku poprawiania działania całego algorytmu.