УДК 004.89

РОЗПІЗНАННЯ МОВИ (ЗВУКУ) ЗА ДОПОМОГОЮ НЕЙРОННИХ МЕРЕЖ

Глімбоцький Владислав

Науковий керівник: канд. ф.-м. наук, доцент Паращук С.Д.

Центральноукраїнський державний педагогічний університет імені Володимира Винниченка, м. Кропивницький, Україна

В статті розглядається питання розробки алгоритмів розпізнання мови за допомогою нейронних мереж. Проаналізовано основні теоретичні положення алгоритми, способи розпізнання мови та розбиття мови на окремі частини, тобто речення, слова, букви . Окрім того, стаття містить детальний опис алгоритму розбиття слів на окремі частини та аналіз отриманого звуку. На його основі можна розробляти алгоритми для інших практичних завдань.

Ключові слова: машинне навчання, розпізнання мови, розпізнання звуків дискретизація.

Recognition of speech using neural networks

V. Hlimbotskyi

Scientific supervisor: Candidate of Physics and Mathematics Sciences, Docent Parashchuk S.D.

The Volodymyr Vynnychenko Central Ukrainian State Pedagogical University, Kropyvnytsky, Ukraine

The article deals with the development of speech recognition algorithms using neural networks. The basic theoretical positions of algorithms, methods of speech recognition and language splitting into separate parts, that is, sentences, words, letters, are analyzed. In addition, the article contains a detailed description of the algorithm for splitting words into separate parts and analysis of the received sound. On its basis, it is possible to develop algorithms for other practical tasks.

Keywords: machine learning, speech recognition, disc sampling sounds recognition.

Постановка проблеми:

Наша мова – це послідовність звуків. Звук в свою чергу – це суперпозиція (накладення) звукових коливань (хвиль) різних частот. Хвилі ж, як відомо з фізики, характеризуються двома атрибутами – амплітудою і частотою (рис.1).

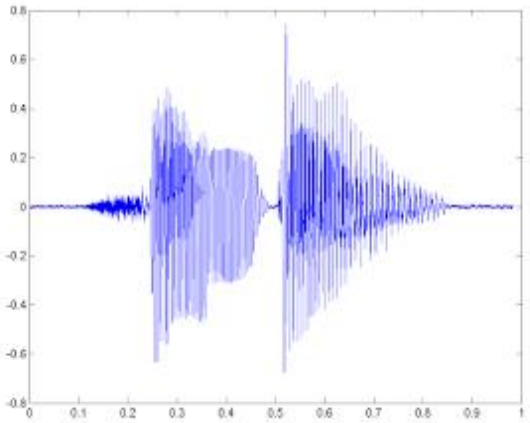

Рис. 1. Характеристики звукової хвилі

Для того, що б зберегти звуковий сигнал на цифровому носії, його необхідно розбити на безліч проміжків і взяти деякий «усереднене» значення на кожному з них (рис.2).

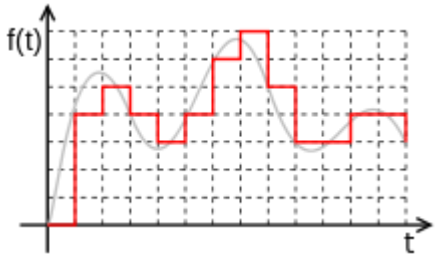

Рис.2. Розбиття звукового сигналу

Таким ось чином механічні коливання перетворюються в набір чисел, придатний для обробки на сучасних ЕОМ. Звідси випливає, що завдання розпізнавання мови зводиться до «порівнянні» безлічі чисельних значень (цифрового сигналу) і слів з деякого словника.

Постановка завдання

Метою статті є висвітлення теоретичних положень отримання, розбиття і розпізнавання мови за допомогою нейронних мереж, а також аналіз основних принципів і алгоритмів обробки звуку.

Вхідні дані

Припустимо у нас є деякий файл/потік з аудіо даними. Перш за все нам потрібно зрозуміти, як він влаштований і як його прочитати. Давайте

розглянемо найпростіший варіант – WAV файл. Формат має на увазі наявність у файлі двох блоків. Перший блок – це заголовок з інформацією про аудіопотоки: бітрейте, частоті, кількості каналів, довжині файлу і т.д. Другий блок складається з «сирих» даних – того самого цифрового сигналу, набору значень амплітуд.

Логіка читання даних в цьому випадку досить проста. Прочитуємо заголовок, перевіряємо деякі обмеження (відсутність стиснення, наприклад), зберігаємо дані в спеціально виділений масив.

Фрейми

Насамперед розіб'ємо наші дані по невеликих тимчасових проміжків - фреймам. Причому фрейми повинні йти не строго один за одним, а один на одному «нероздільно». Тобто кінець одного фрейма повинен перетинатися з початком іншого.

Фрейми є більш придатною одиницею аналізу даних, ніж конкретні значення сигналу, так як аналізувати хвилі набагато зручніше на деякому проміжку, ніж в конкретних точках.

Дослідним шляхом встановлено, що оптимальна довжина фрейму повинна відповідати проміжку в 10мс, «нероздільно» - 50%. З урахуванням того, що середня довжина слова (принаймні в моїх експериментах) становить 500мс - такий крок дасть нам приблизно 500 / (10 * 0.5) = 100 фреймів на слово.

Розбиття слів

Першим завданням, яке доводиться вирішувати при розпізнаванні мови, є розбиття цієї самої мови на окремі слова. Для простоти припустимо, що в нашому випадку мова містить в собі деякі паузи (проміжки тиші), які можна вважати «роздільниками» слів. У такому випадку нам потрібно знайти деяке значення, поріг – значення вище якого є словом, нижче – тишею. Варіантів тут може бути кілька:

- задати константою (спрацює, якщо вихідний сигнал завжди генерується при одних і тих же умовах, одним і тим же способом);

- кластеризувати значення сигналу, явно виділив безліч значень відповідних тиші (спрацює тільки якщо тиша займає значну частину вихідного сигналу);
- проаналізувати ентропію;

Почнемо з того, що ентропія – це міра безладу, «міра невизначеності будь-якого досвіду». У нашому випадку ентропія означає те, як сильно «коливається» наш сигнал в рамках заданого фрейма.

Для того, що б підрахувати ентропію конкретного фрейму слід виконати такі дії:

- припустимо, що наш сигнал пронормувати і все його значення лежать в діапазоні [-1; 1];
- побудуємо гістограму (щільність розподілу) значень сигналу фрейму і : розрахуємо ентропію, як

$$E = \sum_{i=0}^{N-1} P[i] * \log_2(P[i])$$

;

Таким чином, ми отримали значення ентропії. Але це всього лише ще одна характеристика фрейму, і для того, що б відокремити звук від тиші, нам як і раніше потрібно її з чимось порівнювати. У деяких статтях рекомендують брати поріг ентропії рівним середньому між її максимальним і мінімальним значеннями (серед всіх фреймів). Однак, в нашому випадку такий підхід не дав скільки-небудь хороших результатів.

Ентропія (на відміну від того ж середнього квадрата значень) – величина відносно самостійна. Що дозволяє підібрати значення її порога у вигляді константи (0.1). Проте проблеми на цьому не закінчуються: ентропія може просідати по середині слова (на голосних), а може раптово схоплюватися через невеликого шуму. Для того, що б боротися з першою проблемою, доводиться вводити поняття «мінімально відстані між словами» і «склеювати» поблизу лежачі набори фреймів, розділені через просідання. Друга проблема

вирішується використанням «мінімальної довжини слова» і відсіканням всіх кандидатів, які не пройшли відбір (і не використаних в першому пункті).

Якщо ж мова в принципі не є «членороздільною», можна спробувати розбити вихідний набір фреймів на певним чином підготовлені підпослідовності, кожна з яких буде піддана процедурі розпізнавання.

Розкладання в ряд Фур'є

Насамперед розраховуємо спектр сигналу за допомогою дискретного перетворення Фур'є (бажано його «швидкої» FFT реалізацією).

$$X[k] = \sum_{n=0}^{N-1} x[n] * e^{-2*\pi*i*k*n/N}, 0 \leq k < N$$

Так само до отриманими значеннями рекомендується застосувати віконну функцію Хемінга, що б «згладити» значення на кордонах фреймів.

$$H[k] = 0.54 - 0.46 * \cos(2 * \pi * k/(N-1))$$

Тобто результатом буде вектор такого вигляду:

$$X[k] = X[k] * H[k], 0 \leq k < N$$

Важливо розуміти, що після цього перетворення по осі X ми маємо частоту (hz) сигналу, а по осі Y – магнітуду (як спосіб піти від комплексних значень) (Рис.3)

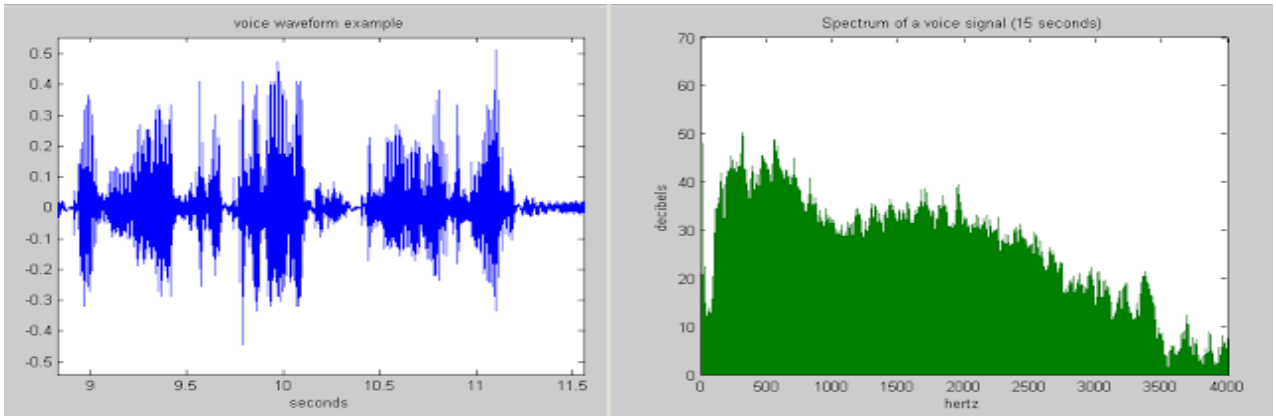

Рис.3. Частота і магнітуда сигналу

Розрахунок mel-фільтрів

Mel - це «психофізична одиниця висоти звуку», заснована на суб'єктивному сприйнятті середньостатистичними людьми. Вона залежить в першу чергу від частоти звуку (а так само від гучності і тембру). Іншими

словами, це величина, що показує, на скільки звук певної частоти «значущий» для нас.

Перетворити частоту в крейда можна за такою формулою:

$$M = 1127 * \log(1 + F/700) \quad (1)$$

Зворотне перетворення виглядає так :

$$F = 700 * (e^{M/1127} - 1) \quad (2)$$

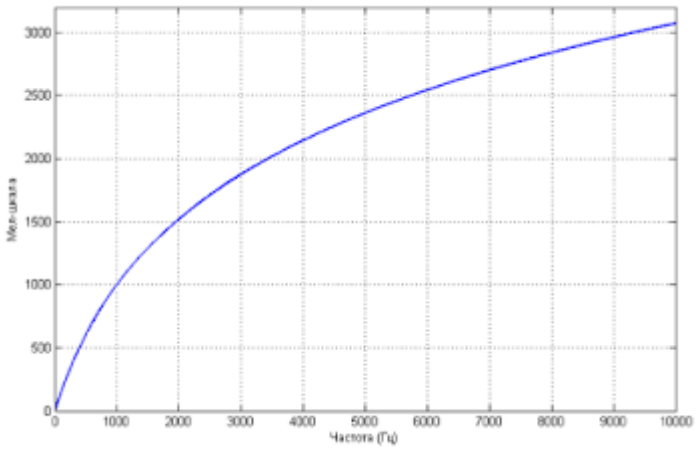

Рис.4. Графік залежності mel / частота

Але повернемося до нашого завдання. Припустимо у нас є фрейм розміром 256 елементів. Ми знаємо (з даних про аудіоформаті), що частота звуку в даній фреймі 16000hz. Припустимо, що людська мова лежить в діапазоні від [300; 8000] hz. Кількість шуканих крейда-коефіцієнтів покладемо M = 10 (рекомендований значення).

Для того, що б розкласти отриманий вище спектр по mel-шкалою, нам буде потрібно створити «гребінку» фільтрів. По суті, кожен mel-фільтр це трикутна віконна функція, яка дозволяє підсумувати кількість енергії на певному діапазоні частот і тим самим отримати mel-коефіцієнт. Знаючи кількість крейда-коефіцієнтів і аналізований діапазон частот ми можемо побудувати набір таких фільтрів (Рис. 5.):

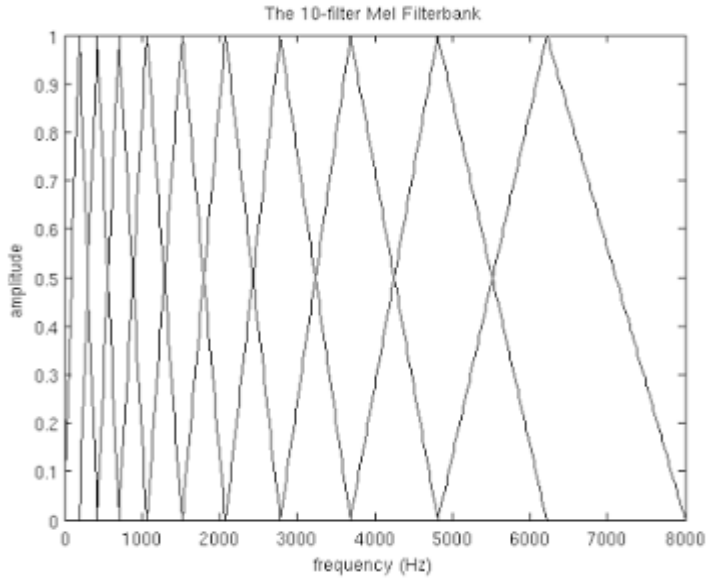

Рис. 5. Набір фільтрів крейд-коефіцієнтів

З рис.5 видно, що чим більше порядковий номер крейда-коефіцієнта, тим ширше основа фільтра. Це пов'язано з тим, що розбиття досліджуваного діапазону частот на оброблювані фільтрами діапазони відбувається на шкалі.

Застосування фільтрів, логарифмування енергії спектра

Застосування фільтру полягає в попарним перемножуванні його значень зі значеннями спектра. Результатом цієї операції є mel-коефіцієнт. Оскільки фільтрів у нас M, коефіцієнтів буде стільки ж.

$$S[m] = \log(\sum_{k=0}^{N-1} |X[k]|^2 * H_m[k]), 0 \leq m < M$$

Однак, нам потрібно застосувати mel-фільтри не до значень спектра, а до його енергії. Після чого прологаріфмувати отримані результати. Вважається, що таким чином знижується чутливість коефіцієнтів до шумів.

Косинусне перетворення

Дискретне косинусне перетворення (DCT) використовується для того, що б отримати ті самі «кепстральних» коефіцієнти. Сенс його в тому, що б «стиснути» отримані результати, підвищивши значимість перших коефіцієнтів і зменшивши значимість останніх.

В даному випадку використовується DCTII без будь-яких добутків на scale factor.

$$C[l] = \sum_{m=0}^{M-1} S[m] * \cos(\pi * l * (m + \frac{1}{2})/M)), 0 \leq l < M$$

Тепер для кожного фрейма ми маємо набір з M mfcc-коефіцієнтів, які можуть бути використані для подальшого аналізу.

Висновки

В ході дослідження було проаналізовано процес отримання вхідних аудіо даних, фрейми звуку і їх відмінність. Наведені способи розбиття і аналізу мови. Розглянуті спосіб розбиття мови за допомогою рядів Фурье, а також застосування і розрахунок mel-фільтрів.

Список літератури

1. Нейронные сети для распознавания речи [Електронний ресурс]/ Дмитриев Е.А. – Режим доступу до ресурсу: https://interactive-plus.ru/ru/article/463478/discussion_platform

2. Компьютерное распознавание и порождение речи. [Електронний ресурс]. – Режим доступу до ресурсу: http://speech-text.narod.ru/chap3.html

3. Маркова, В.А. Сети Джордана и Элмана. [Електронний ресурс]. – Режим доступу до ресурсу: http://i-intellect.ru/articles-of-neural-networks/jordans-and-elmans-networks.html

4. Стариков, А. Генетические алгоритмы – математический аппарат. [Електронний ресурс]. – Режим доступу до ресурсу: http://www.basegroup.ru/library/optimization/ga_math/

5. Распознавание команд с помощью ДПФ и библиотеки нейронной сети FANN [Електронний ресурс]/ Мясищев А.А. - Режим доступу до ресурсу: https://sites.google.com/site/webstm32/raspoznavanie-komand-dpf-fann